IEEE Signal Processing MAGAZINE
Volume 34 | Number 4 | July 2017

HIGH-IMPACT INNOVATIONS

Powered by Signal Processing

Geometric Deep Learning

Adaptive Importance Sampling

Image Aesthetic Assessment

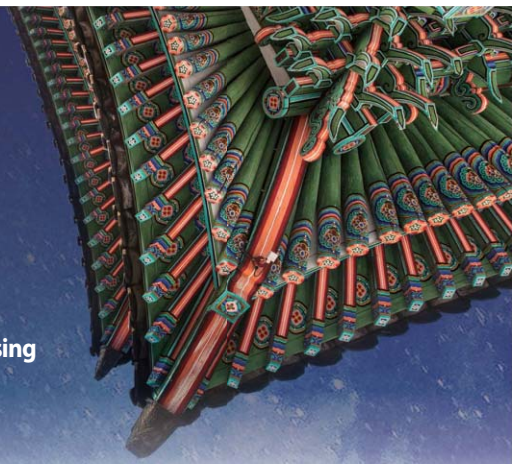
Beat Detection Meets Embedded System at the SP Cup

IEEE Signal Processing Society

IEEE

Call for Papers and Sponsors

ICASSP2018



The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing

April 22 - 27, 2018, Seoul, Korea

http://2018.ieeeicassp.org

Signal Processing and Artificial Intelligence: Changing the World

Submission of Papers

Authors are invited to submit papers of not more than four pages of technical content including figures and references, with an optional fifth page containing only references. Submission instructions, paper format templates, and other important information will be made available on the ICASSP 2018 website, http://2018.ieeeicassp.org.

Conference Topics

The conference will feature world-class international speakers, tutorials, exhibits, lectures and poster sessions from around the world. Topics include but are not limited to:

- Audio and acoustic signal processing
- Sensor array & multichannel signal processing
- Bio-imaging and biomedical signal processing
- Signal processing education
- Design & implementation of signal processing systems
- Signal processing for communications & networking
- Image, video & multidimensional signal processing
- Signal processing theory & methods
- Industry technology tracks
- Signal processing for big data
- Information forensics and security
- The Internet of Things & RFID
- Machine learning for signal processing
- Speech processing
- Spoken language processing
- Multimedia signal processing
- Remote sensing and signal processing
- Signal processing for brain machine interface
- Signal processing for smart systems
- Signal processing for cyber security
- Computational imaging

Call for Tutorials

Tutorials at ICASSP form an important part of the program, giving attendees the opportunity to learn about current research areas that are of growing interest to the signal processing community. Those who are interested in presenting a tutorial may want to contact one of the tutorial chairs before preparing a formal proposal. It is important to keep in mind, for any tutorial, that it should be tutorial in nature, and within the grasp of a wide audience.

Call for Special Sessions

The program for ICASSP 2018 will include Special Sessions that complement the traditional program with new and emerging topics of significant interest to the signal-processing community, particularly those that are in line with the theme of the conference. Please refer to the conference webpage for information about Special Session proposals.

Call for Exhibitors and Sponsors

ICASSP 2018 offers exhibitors and sponsors an opportunity to showcase their company's products and innovative solutions at the Signal Processing Society's flagship conference that will be held for the first time in the Korean Peninsula. Please refer to the conference webpage for information about signing up to become an exhibitor or sponsor at ICASSP.

Signal Processing Letters

Authors of IEEE Signal Processing Letters (SPL) papers will be given the opportunity to present their work at ICASSP 2018, subject to space availability and approval by the Technical Program Chairs. SPL papers published between January 1, 2017 and December 31, 2017 are eligible for presentation at ICASSP 2018. Because they are already peer-reviewed and published, SPL papers presented at ICASSP 2018 will neither be reviewed nor included in the proceedings.

Important Dates

August 4, 2017
Special Session Proposals Due

August 11, 2017
Tutorial Proposals Due

September 8, 2017
Notification of Special Session Acceptance

September 15, 2017
Notification of Tutorial Acceptance

October 27, 2017
Paper Submissions Due

January 12, 2018
Signal Processing Letters Due

January 26, 2018
Notification of Paper Acceptance

February 9, 2018
Revised Paper Upload Deadline

February 16, 2018
Author Registration Deadline

General Chairs

Monson Hayes
Hanseok Ko

Technical Program Chairs

Dan Schonfeld
Pascale Fung
Nam Ik Cho

Sponsored by

IEEE

IEEE Signal Processing Society

Digital Object Identifier 10.1109/MSP.2017.2710839

Contents

Volume 34 | Number 4 | July 2017

FEATURES

Theory and Methods

18 GEOMETRIC DEEP LEARNING
Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst

43 OPTIMAL MASS TRANSPORT
Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepčev, and Gustavo K. Rohde

60 ADAPTIVE IMPORTANCE SAMPLING
Mónica F. Bugallo, Víctor Elvira, Luca Martino, David Luengo, Joaquín Míguez, and Petar M. Djurić

ON THE COVER

This issue of *IEEE Signal Processing Magazine* showcases high-impact innovations in signal processing as well as recent advances and activities through editorials, feature articles, and a diverse set of columns.

COVER IMAGE: ©ISTOCKPHOTO.COM/MONSITJ

Visual and Speech Analytic

80 IMAGE AESTHETIC ASSESSMENT
Yubin Deng, Chen Change Loy, and Xiaoou Tang

107 ADVANCED DATA EXPLOITATION IN SPEECH ANALYSIS
Zixing Zhang, Nicholas Cummins, and Björn Schuller

Emerging Applications

130 DRIVER MODELING FOR DETECTION AND ASSESSMENT OF DISTRACTION
John H.L. Hansen, Carlos Busso, Yang Zheng, and Amardeep Sathyanarayana

COLUMNS

6 Reader's Choice
Top Downloads in IEEE *Xplore*

8 Society News
Nominations Open for 2017 IEEE Signal Processing Society Awards

10 Community Voices
What Do You Consider a "Successful" Career?
Andres Kwasinski and Min Wu

14 Special Reports
Innovative Sensors Promise Longer and Healthier Lives
John Edwards

143 SP Competitions
Embedded Systems Feel the Beat in New Orleans
Craig T. Jin, Matthew E.P. Davies, and Patrizio Campisi

151 SP Education
A Raspberry Pi-Based Platform for Signal Processing Education
Gianni Pasolini, Alessandro Bazzi, and Flavio Zabini

On "Flipping" a Large Signal Processing Class
Waheed U. Bajwa

171 Lecture Notes
Demystifying Compressive Sensing
Heinrich Edgar Arnold Laue

Vertex-Frequency Analysis: A Way to Localize Graph Spectral Components
Ljubiša Stanković, Miloš Daković, and Ervin Sejdić

183 Tips & Tricks
Digital Envelope Detection: The Good, the Bad, and the Ugly
Richard Lyons

188 Conference Highlights
It Really Was Lagniappe!
Magdy Bayoumi

196 In the Spotlight
How Biometric Authentication Poses New Challenges to Our Security and Privacy
Nasir Memon

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US$20.00 (first copy only), nonmembers US$241.00 per copy. Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 **Printed in the U.S.A.**

Digital Object Identifier 10.1109/MSP.2017.2693141

DEPARTMENTS

3 **From the Editor**
Innovations Powered by Signal Processing
Min Wu

4 **President's Message**
Mind the (Gender) Gap
Rabab Ward

192 **Dates Ahead**

Digital Object Identifier 10.1109/MSP.2017.2693143

SCOPE: *IEEE Signal Processing Magazine* publishes tutorial-style articles on signal processing research and applications as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.

Min Wu | Editor-in-Chief | minwu@umd.edu

Innovations Powered by Signal Processing

To many people around the world, Abraham Lincoln was a highly regarded president of the United States whose pursuit of social justice paved the way to end the slavery in this country. Fewer people, however, know his distinction as an engineering innovator and that he was the first and only U.S. president thus far to hold a patent. President Lincoln received the U.S. Patent 6469 in May 1849 titled "Buoying Vessels over Shoals," which was inspired by his experience navigating boats on the Ohio and Mississippi Rivers. When commenting on the role of the patent system that offers the inventor the exclusive use of his/her invention for a limited time, President Lincoln noted that the patent system "added the fuel of interest to the fire of genius, in the discovery and production of new and useful things."

From about a century after Lincoln's invention and throughout the next several decades, signal processing has contributed significantly to technology innovations and changed how we work and live. Smartphones, digital photography, the global positioning system, and medical diagnosis are tangible examples around us, and our magazine has touched on them through special issues and column articles. What many of us take for granted and wouldn't pause to think about are numerous examples that we don't see, such as when we store our data on hard drives. This is one of the best examples that recently comes to mind to showcase the profound impact of signal processing research.

Aleksandar Kavcic and José Moura's academic research at Carnegie Mellon University in the 1990s studied the effect when data would become densely packed in magnetic disk drives, and they proposed signal processing algorithms to enable the accurate detection of data stored in high-density disks—which became the norm a decade later in billions of computers. Their pioneering research also made front-page news when a US$750 million settlement was announced concerning the infringement of their corresponding patents, the second-largest payment over any technology patents to date. You can read more about the Kavcic–Moura detector for high-density magnetic recording on the IEEE Signal Processing Society's online blog: http://signalprocessingsociety.org/publications-resources/blog/why-signal-processing-pioneer-takes-road-less-traveled.

Innovations have continuously come from both industry and universities, often in complementary ways, although there may be stereotypical views on the roles that each side plays. Every once in a while, I have friends working in industry questioning the practical values of university research beyond writing papers and training students. Indeed, many publications may not see widespread real-world use. More often, we see academic publications as well as industrial products making incremental improvements over the prior art as opposed to making revolutionary advances, and it is common that exploratory research has been carried out well before the market ecosystem or the supporting technologies to become ready.

The immediacy of deployment is perhaps one of the differentiating factors between product development and exploratory research, but as history reveals, it is not the primary indicator of the impact of innovations. Kavcic and Moura showed to the world the impact of their innovations from what started out as exploratory research in signal processing. One attribute that enables their impact (as opposed to a purely intellectual exercise or a bean-counting effort to add to one's publication list) is the educated anticipation of the technological trends (in their case, the increasing density of the data being packed in storage drives) and the willingness to tackle challenges beyond making epsilon-delta improvement.

To qualify my words here, incremental improvements have their important roles in technological advances, and many progresses made—big or small—are standing on the shoulders of giants; we continuously build on the efforts of our technical community in direct and indirect ways. But the willingness to go beyond, to nurture out-of-the-box thinking, and to encourage taking higher risks opens up opportunities for bigger leaps in innovation, even if we may not succeed at most attempts.

(continued on page 9)

Digital Object Identifier 10.1109/MSP.2017.2705258
Date of publication: 11 July 2017

PRESIDENT'S MESSAGE

Rabab Ward | SPS President | rababw@ece.ubc.ca

Mind the (Gender) Gap

In my last "President's Message" [1], I talked about the many ways that diversity drives innovation in our field and in all facets of life. Returning to this discussion, I'd now like to focus on women in engineering (WIE). There are numerous examples of women throughout history who had to overcome serious hurdles to make valuable contributions to science and culture. I will mention three exemplary role models.

Sophie Germain, the great French mathematician, was not allowed to attend the Ecole Polytechnique in the late 1700s; she used a male pseudonym and eventually won the Paris Academy of Science grand prize for the theory of elasticity. Another famous renegade in popular culture and in the signal processing field is the actress Hedy Lamarr, who invented a spread spectrum technology in 1941 to scramble radio signals on torpedoes. In the 1960s, her technology was implemented on naval ships during the Cuban Missile Crisis. Her applications ultimately galvanized the digital communications boom and are currently used in many devices, including the global positioning system, cell phones, and fax machines.

Last but not least is Madam Curie, whose incredible achievements influenced me from an early age. She was my idol, and I used to dream that one day I would also do great things. As a young girl, I did not know exactly what my own "great thing" would be. But I was lucky to also have a mother who instilled in me the confidence that, with hard work, I could reach any goals I set. I certainly faced many obstacles, but whenever I came up against a closed door, I looked for another door to open.

I was born and raised in Lebanon. In 1961, I was not allowed to study engineering at the American University of Beirut, even though my grades were higher than all of the male students in the country at that time. So I went to Egypt and enrolled in medical school. But my heart wasn't in medicine, so I switched to engineering. I returned home in 1967 and became the first female member of the Lebanese Professional Engineering Society. When I later joined the University of California in Berkeley, I was the only woman among all of the Ph.D students in electrical engineering. In fact, the first woman to earn her Ph.D. degree in electrical engineering at Berkeley was an Egyptian woman who graduated four years before me.

Upon my graduation in 1972, unlike my male colleagues, I could not find a job in academia. I looked upon that as an opportunity—to have children and also to work abroad, which led to my becoming the first woman appointed in engineering at the University of Zimbabwe. Eventually I joined the University of British Columbia's Faculty of Applied Science, and, in 1981, I received a tenure track professorship and became the first woman engineer professor at British Columbia.

Looking back, I realize that the vast majority of those who helped me succeed were males as, at that time, I did not know other female engineers in my field who could extend a mentoring hand. I've enjoyed a fulfilling career in electrical engineering, specifically in the signal processing field. And I've since enjoyed the privilege of working with many talented women students, colleagues, scientists, engineers, innovators, and business people.

The number of women studying and working in science, technology, engineering, and mathematics (STEM) has increased so much since I was a girl. But since the early 2000s, the number of women engineering undergraduates has remained stagnant at about 20% in the United States, and only 11% of practicing engineers in the United States are women, with typically lower annual salaries lower than men (in 2013, it was US$65,000 for women compared to US$79,000 for men). It's no wonder that only 27% of women remain in the STEM fields after the age of 30.

According to 2016 research, STEM Fortune 500 firms are no more diverse than in 2001, indicating an entrenched gender gap. The picture is similar in the United Kingdom, where only 9% of the engineering workforce are women. At the same time, the United States and the United Kingdom are dealing with a shortage of STEM workers. With STEM jobs set to grow 17% by 2024, we need to provide more resources and incentives to women, to fill the labor market gap. Furthermore, according to research, companies with gender parity are 15% more likely to perform better.

There is encouraging news coming from other countries. Thirty-five percent of engineering students in India are women. In some Arab countries, enrollments are

Digital Object Identifier 10.1109/MSP.2017.2704998
Date of publication: 11 July 2017

as high as 60%. The number of women studying and working in computer science has surged, and by 2014, 25% of startups in the Middle East were owned by women. In China, women account for approximately 40% of the STEM workforce. However, women in every country have to contend with their own culture's biases and workplace challenges.

At the IEEE, the percentage of female Members is approximately 10.6%—a 3% point increase (from 7.5%) since 2000. Over the past few years, the IEEE WIE has become one of the world's largest international professional organizations, dedicated to the professional development of women. WIE has made big strides, growing from 3,000 members in 2001 to 15,000 in 2013 and is now extremely active in more than 70 countries.

During an outreach activity for the SPS in 2015, I met an outstanding female engineer, Maryam Al Thani, in Abu Dhabi who is very active with WIE. That year she ran for election to the United Arab Emirates Federal National Council (the body that represents the Emirates' nationals). How how did she acquire the confidence, at this relatively young age, to run for the highest position in her country? She told me that her leadership skills came by actively volunteering in IEEE WIE. Although she did not win that election, she is confident that she will one day.

Today, women account for 9.4% of the IEEE Signal Processing Society's (SPS's) membership. The SPS started holding the Women in Signal Processing (WISP) luncheon at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 1997. Since 2015, this luncheon has been held at the majority of our large conferences: ICASSP, the IEEE International Conference on Image Processing, and GlobalSIP. Each luncheon features an invited guest speaker, discussions, and networking. Although it is called Women in Signal Processing, men also enjoy attending this event.

This year, the activities of the SPS committee on women have expanded beyond conference events. During the WISP Luncheon at ICASSP 2017, we debuted the new Women in Signal Processing Directory. The directory, visualized by Namrata Vaswani, will act as a resource for women in the SPS and engineering, building a global community of women in signal processing fields and positioning them to gain visibility and raise awareness about opportunities for leadership roles, award nominations, and more.

A recent report on women in the SPS by a committee chaired by Mari Ostendorf found that women make up approximately 15% of our technical committees, 10.7% of our associate editors, and 17% of leadership roles in our Society. The percentage of women winning major SPS (nonservice) awards since 1990 (2.2%) is, however, much lower than the percentage of female fellows (10%). We hope that the establishment of the Women in Signal Processing Directory will increase visibility of women in the field to be considered for awards and leadership roles.

There is a delicate line to walk between inclusivity and tokenism. How can we include and empower women without exploiting their gender or placing too much focus on gender? I think it starts with not only nurturing their confidence but having confidence in them and their abilities. The availability of hands-on experience in elementary and high schools is crucial as is the dedication of teachers who truly believe in girls' abilities, challenging and encouraging boys and girls equally. A supportive, inclusive network would provide girls and women with the tools to build communities that motivate them to persist in this field.

Gender inclusivity in engineering is not only good for society, it is also beneficial for business. It fuels innovation and enriches every facet of our life. Let's do more to make our businesses, our research labs, our academic institutions, and our domestic environments enriching places, where girls and women can thrive.

Reference

[1] R. Ward, "Diversity through adversity," *IEEE Signal Process. Mag.*, vol. 34, no. 3, pp. 4–5, May 2017.

SP

READER'S CHOICE

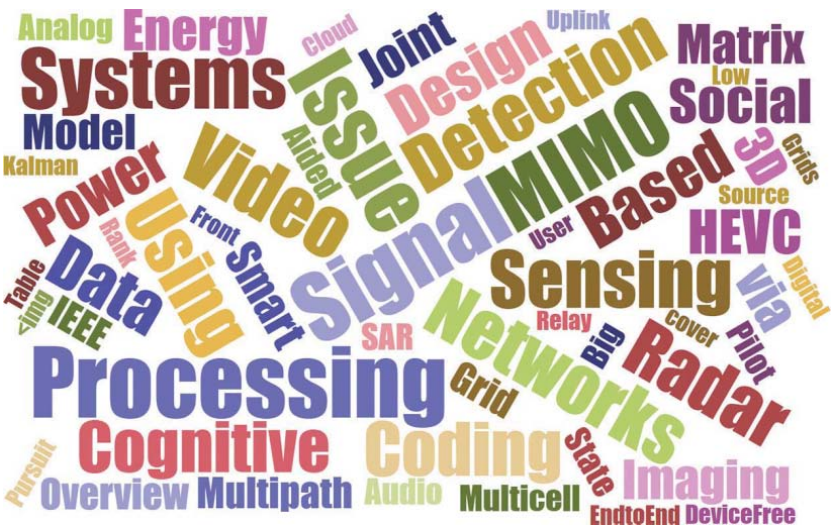
Top Downloads in IEEE *Xplore*

Each "Reader's Choice" column in *IEEE Signal Processing Magazine* focuses on a different publication of the IEEE Signal Processing Society (SPS). This issue of the magazine highlights articles in *IEEE Journal of Selected Topics in Signal Processing* (*JSTSP*).

JSTSP solicits special issues on topics that cover the entire scope of the SPS. *JSTSP* only publishes papers that are submitted in response to a specific call for papers, which can be found at https://signalprocessingsociety.org/publications-resources/ieee-journal-selected-topics-signal-processing. Instructions for submitting papers to a particular special issue can be found by clicking on the "Information for Authors" tab under the "*JSTP* Menu" on this page. The procedure for preparing and submitting a proposal for a special issue can be found by selecting the "Special Issue Proposal Submission" tab. All special issue proposals are evaluated by our senior editorial board for relevance, timeliness, technical merit, impact, and general interest to the Society.

This issue's "Reader's Choice" column lists the top ten papers most downloaded from January 2015 to March 2017. Your suggestions and comments are welcome and should be sent to Associate Editor Chungshui Zhang at zcs@mail.tsinghua.edu.cn.

Digital Object Identifier 10.1109/MSP.2017.2697178
Date of publication: 11 July 2017



An Overview of Massive MIMO: Benefits and Challenges

Lu, L.; Li, G.Y.; Swindlehurst, A.L.; Ashikhmin, A.; Zhang, R.

In this paper, the authors present a comprehensive overview of state-of-the-art research on the topic. The paper includes an informational theoretic analysis; implementation issues related to channel estimation, detection, and precoding schemes; the energy efficiency; and the challenges and opportunities.

October 2014

Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems

Alkhateeb, A.; Ayach, O.E.; Leus, G.; Heath, R.W.

This paper develops an adaptive algorithm to estimate the millimeter wave channel parameters, which exploits the poor scattering nature of the channel. Furthermore, a novel hierarchical multiresolution codebook is designed to construct training beamforming vectors with different beamwidths.

October 2014

Standardized Extensions of High Efficiency Video Coding (HEVC)

Sullivan, G.J.; Boyce, J.M.; Chen, Y.; Ohm, J.-R.; Segall, C.A.; Vetro, A.

This paper describes extensions to the High-Efficiency Video Coding (HEVC) standard that are active areas of current development in the relevant international standardization committees. The design

©GRAPHICSTOCK

for these extensions represents the latest state of the art for video coding and its applications.

December 2013

MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio

Herre, J.; Hilpert, J.; Kuntz, A.; Plogsties, J.

The ISO/MPEG standardization group has started the MPEG-H 3D Audio development effort to facilitate high-quality bit rate-efficient production, transmission, and reproduction of such immersive audio material. This paper provides an overview of the MPEG-H 3D Audio project and technology and an assessment of the system capabilities and performance.

August 2015

Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems

Figueiredo, M.A.T.; Nowak, R.D.; Wright, S.J.

This paper proposes gradient projection (GP) algorithms for the bound-constrained quadratic programming formulation of sparse reconstruction. Computational experiments show that these GP approaches perform well in a wide range of applications, often being significantly faster (in terms of computation time) than competing methods.

December 2007

Advances in Cognitive Radio Networks: A Survey

Wang, B.; Liu, K.J.R.

This paper surveys recent advances in research related to cognitive radios. The fundamentals of cognitive radio technology and architecture of a cognitive radio network and its applications are introduced. The existing works in spectrum sensing are reviewed, and important issues in dynamic spectrum allocation and sharing are investigated in detail.

February 2011

An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems

Heath, R.W.; González-Prelcic, N.; Rangan, S.; Roh, W.; Sayeed, A.M.

This article provides an overview of signal processing challenges in millimeter wave wireless systems, with an emphasis on those faced by using multiple-input, multiple output communication at higher carrier frequencies.

April 2016

A Real-Time End-to-End Multilingual Speech Recognition Architecture

Gonzalez-Dominguez, J.; Eustis, D.; Lopez-Moreno, I.; Senior, A.; Beaufays, F.; Moreno, P.J.

In this paper, the authors present an end-to-end multilingual automatic speech recognition system architecture, developed and deployed at Google, that allows users to select arbitrary combinations of spoken languages. They leverage recent advances in language identification and a novel method of real-time language selection to achieve similar recognition accuracy and nearly identical latency characteristics as a monolingual system.

June 2015

Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays

Sohrabi, F.; Yu, W.

This paper considers a hybrid beamforming architecture in which the overall beamformer consists of a low-dimensional digital beamformer followed by a radio-frequency (RF) beamformer implemented using analog phase shifters. This paper establishes that if the number of RF chains is twice the total number of data streams, the hybrid beamforming structure can realize any fully digital beamformer exactly, regardless of the number of antenna elements.

April 2016

A Survey of Stochastic Simulation and Optimization Methods in Signal Processing

Pereyra, M.; Schniter, P.; Chouzenoux, É.; Pesquet, J.-C.; Tourneret, J.-Y.; Hero, A.O.; McLaughlin S.

This survey paper offers an introduction to stochastic simulation and optimization methods in signal and image processing. The paper addresses a variety of high-dimensional Markov chain Monte Carlo methods as well as deterministic surrogate methods, such as variational Bayes, the Bethe approach, belief and expectation propagation, and approximate message-passing algorithms.

March 2016

SP

Nominations Open for 2017 IEEE Signal Processing Society Awards

The IEEE Signal Processing Society (SPS) Awards Board is now accepting nominations for all Society-level awards, from paper awards to the major society awards. Nominations are due by 1 September 2017 and should be submitted to Theresa Argiropoulos (t.argiropoulos@ieee.org), who will collect the nominations on behalf of Awards Board Chair H. Vincent Poor. Nominators should take into consideration the need for representation of diversity in the nomination slate when submitting their nominations. Detailed information and nomination/endorsement forms for SPS awards can be found online. Full details on the nomination process are available at http://signalprocessingsociety.org/get-involved/awards-submit-award-nomination.

Please note that, this year, the Society will be testing new awards software, so nominations for the Technical Achievement Award must be submitted online through this link: https://ieee.secure-platform.com/a/page/society_awards/ieeesignalprocessingsocietyawards.

All other awards will be handled though the normal submission process and should be submitted to Theresa Argiropoulos via e-mail.

- *Who can nominate*: Nominations are accepted from any Society individual member, Society committee, or Society board. Nominations from individual members can be supported by up to two endorsement forms from two other individual members.
- *Which Awards*: Each year, the SPS honors outstanding individuals who have made significant contributions related to signal processing through the Society Award, the Industrial Leader Award, the Industrial Innovation Award, the Technical Achievement Award, the Education Award, the Meritorious Service Award, and the Meritorious Chapter/Regional Service Award. The Society also recognizes outstanding publications in SPS journals and the magazine through the Best Paper Award, Donald G. Fink Overview Paper Award, Sustained Impact Paper Award, Signal Processing Letters Best Paper Award, Signal Processing Magazine Best Column Award, Signal Processing Magazine Best Paper Award, and the Young Author Best Paper Award.

Nominations for the Best Paper Award and Young Author Best Paper Award should refer to the papers published in the following Society journals:

- *IEEE Journal of Selected Topics in Signal Processing* (*JSTSP*)
- *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (*T-ASLP*)
- *IEEE Transactions on Image Processing* (*T-IP*)
- *IEEE Transactions on Information Forensics and Security* (*T-IFS*)
- *IEEE Transactions on Signal Processing* (*T-SP*).

SPS awards changes

Over the past few years, the Society approved some procedural changes to the SPS Awards program, including some new changes approved late last year. Please note that these changes are in effect for the 2017 nomination period. The changes are intended to provide an effective means to encourage award nominations in all categories from the SPS community at large, including individuals, technical committees, editorial boards, and other major boards, except in the cases of conflict of interests. Technical committees and boards may pass on to the Awards Board one or multiple nominations that they receive for all awards.

The Society created a new award called the Meritorious Regional/Chapter Service Award, which focuses on outstanding contributions of any member of the Society to regional activities of the SPS. As a result of the creation of this award, the judging criteria for the Meritorious Service Award was redefined. The Meritorious Service Award judging criteria now reflects that judging will be based on recognizing outstanding efforts and contributions aimed at promoting the technical and educational activities of the entire SPS, i.e., that benefit the membership of the SPS at large.

Digital Object Identifier 10.1109/MSP.2017.2697105
Date of publication: 11 July 2017

The Overview Paper Award was renamed the Donald G. Fink Overview Paper Award. The award description remains untouched; just the name of the award was modified.

The Society Award was modified to incorporate a presentation of an International Conference on Acoustics, Speech, and Signal Processing (ICASSP) plenary lecture, which will be called the "Norbert Wiener Lecture." Each Society Award recipient is expected to present a Norbert Wiener lecture at the 2018 ICASSP. This lecture is one of the plenary lectures given on the day of the banquet of ICASSP, but it is not a banquet speech.

As a reminder, for the Young Author Best Paper Award, a board or committee cannot nominate one of its members for the award. Please note that this includes nominating an author of a paper where a member of a nominating board or committee is also an author on the paper, even though this member is not the "young author" being considered for the award.

The paper awards nomination form requests citation impact information, so please provide this valuable information. The Awards Board will continue to review the nominations and make selections on paper awards.

For all major awards other than paper awards, the Awards Board will be responsible for vetting the nominations and producing a short list of no more than three nominations per award. The Board of Governors will continue to vote on the selection of the major awards.

A board or committee cannot nominate one of its current members for an award. However, the board/committee member can be nominated by another board or committee. Current elected members of a committee/board may participate as individual nominators for other members of the same board/committee. In the case of major award nominations, please note: boards or committees that submit nominations, but have voting Board of Governors members sitting on their boards or committees, must ensure that Board of Governors voting members do not participate in the board/committee award nomination or selection process.

Individual nominations can have multiple conominators listed on the nomination form. In addition, individual nominations can include up to two endorsements to strengthen the nomination from two other individual members. Nominations supported by committee/boards cannot be accompanied by endorsements. IEEE SPS membership is no longer required for endorsements. All endorsements must be submitted via e-mail to the specified address, which will provide the nomination with a date and time stamp. If more than two endorsements are submitted, only the first two received endorsements will be forwarded to the SPS Awards Board for consideration. A nominator cannot serve as an endorser for a nomination he/she is submitting. If the Society policies state that a particular board/committee/individual is not eligible to nominate for a particular award, then members of that same group of individuals are not eligible to be endorsers.

Technical committee and special interest group award nomination procedures have been approved with suggested award nomination and voting procedures. For full details on each award as well as the new Society and technical committee/special interest group awards policies and endorsement form, please visit http://signalprocessingsociety.org/get-involved/awards-submit-award-nomination.

If you have any questions regarding the process, please do not hesitate to contact Awards Board Chair H. Vincent Poor at poor@princeton.edu.

SP

FROM THE EDITOR

(continued from page 3)

opens up opportunities for bigger leaps in innovation, even if we may not succeed at most attempts.

As in almost any litigation, for colleagues who either work for or hold shares and other interests in the opposing company involved in the patent dispute, the success of the inventors and their institution in this high-profile litigation may be rather bitter. This is understandable as one's judgment can be influenced when such personal interests are involved. Still, I hope as professionals working on the forefront of technology advances, we can look beyond our personal gains or losses to celebrate the positive impact of innovations powered by signal processing.

Perhaps our discussions on the innovations powered by signal processing have stimulated reflections from you. To help capture the thoughts of our readers, we formally launch the "Community Voices" column on page 10 in this issue. The first discussion topic is "What is considered a successful career for signal processing trained professionals?" *SPM*'s Area Editor Dr. Andres Kwasinski took the lead and gathered input from the community and compiled highlights. My appreciation also goes to Dr. Charles Casimiro Cavalcante, a reader in Brazil, who was the very first to respond to the open calls on this new initiative, and to several readers from a variety of sectors together with our retired veterans of the magazine editorial board who kindly share their perspectives.

The second topic for the "Community Voices" column is "What's the future of signal processing?" Please take a moment to share your views on this web form https://www.surveymonkey.com/r/SPSCommunityVoices2. We look forward to reading your input and sharing highlights in a future issue of *SPM*.

SP

Andres Kwasinski and Min Wu

What Do You Consider a "Successful" Career?

Perspectives from signal processing-trained professionals

Welcome to the first article in a newly launched column, "Community Voices," in *IEEE Signal Processing Magazine*. The motivation behind this column is to strengthen ties with readers and members in the signal processing community. In doing so, we set out to collect reflections from diverse members of our community on questions that are of interest to many. A readily available form on the Internet as well as e-mail exchanges were used to gather responses. This first article of the "Community Voices" column focuses on the question "For a person with signal processing training, what do you consider as a successful career?"

We begin with input from Charles Casimiro Cavalacante from Brazil, who was the first to respond to the web form. We welcome your feedback on this new initiative and your ideas in suggesting future topic questions. The second topic on the future of signal processing is open for input. Please refer to "The Future of Signal Processing" for the topic and web links. We hope that you enjoy this new column and look forward to hearing from you.

Charles Casimiro Cavalacante

Signal processing is broad, and career prospects for signal processing practitioners are just as diverse. There are practitioners in biomedical engineering, industrial automation, electronic design, acoustics and audio applications, image and video processing; robotics, navigation systems, data and financial analytics, communication systems, and many others. There are growing research areas in data analytics, perceptual computing, smart energy technologies, and sensor systems for enterprise and industrial applications. There is a wealth of signal processing expertise in research institutions pushing signal processing reach into many fields through research projects and training of the next cadre of practitioners.

While judging career success is a subjective exercise, there are good indicators common to most people's ideal of a successful career. These include drawing satisfaction from day to day job activities, progressive growth in project responsibilities and influence, and doing recognized and rewarding work that has measurable impact.

Given the breadth of signal processing career opportunities and understanding what constitutes career success, what does a successful signal processing career look like? I am a midcareer practitioner with experience in both academia and in industry. I consider myself a work in progress toward career success. Signal processing has enabled me to contribute to modeling high-speed computer interconnects and gain insight into channel equalization challenges, train students on filter design, and witness the excitement of translating design-rule steps to circuit implementation for a rudimentary working guitar pickup. These are some rungs on a ladder toward a satisfying career.

Listening to senior engineers discuss their most impactful work and the

The Future of Signal Processing

After half a century of development, some say signal processing is already matured in terms of theories and techniques and perhaps would not have a new research breakthrough. Others have observed the problem of "signal processing inside."

What are your thoughts about the future of signal processing? Please provide your input by filling out this web form: https://www.surveymonkey.com/r/SPSCommunityVoices2.

A selection of the responses will be published in an upcoming issue of *IEEE Signal Processing Magazine* or *Inside Signal Processing eNewsletter,* subject to editing for language and length.

Digital Object Identifier 10.1109/MSP.2017.2698118
Date of publication: 11 July 2017

process that took them from ideation to results shows that careers are a journey indeed. In short, a successful career is a hodgepodge of experiences, growth through overcoming challenges, project successes, and willingness to embrace new ways of using signal processing training in different engineering problems.

Author

Charles Casimiro Cavalcante (charles@gtel.ufc.br) received a Ph.D. degree from the University of Campinas. He is an associate professor at the Universidade Federal do Ceará, Brazil, and holds the Statistical Signal Processing Chair. He has been a visiting assistant professor in the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. He is a Senior Member of the IEEE and of the Brazilian Telecommunications Society.

Khaled El-Maleh

I have witnessed the great impact of signal processing in our lives! With signal processing training spanning three degrees (my B.Eng., M.Eng., and Ph.D. degrees) followed by more than 17 years of industry experience, I have been fortunate to have a successful and rich signal processing career. I think the main reasons for this success have been realizing the great value of continuing my relationship with academic institutions (both students and professors), with the IEEE Signal Processing Society, as well as working on developing multimedia consumer products using signal processing algorithms. Examples of such products are smartphones with wide-band telephony, advanced camera and video telephony, and streaming. In addition, I have recently expanded my signal processing knowledge in emerging areas like the Internet of Things, automotive, mobile health care, and smart cities.

Author

Khaled El-Maleh (kelmaleh@qualcomm.com) received his Ph.D. degree in electrical and computer engineering from McGill University, Canada. He is a senior director of technology in the Intellectual Property (IP) Department of Qualcomm leading the Sensor and Display IP Portfolio Team, Multimedia Technology Team, and related IP Strategy areas. He is a technologist and strategist with focus on entrepreneurship and innovation, as well as an accomplished inventor with more than 200 U.S. and international patents. He was awarded the Qualcomm Career Thought Leadership Award in 2009 and the IP Department 2013 Distinguished Contributor Award.

Gene A. Frantz

I first will start with an equation, which is the basis for my answer: DSP + Divide = Math.

When we began the drive into digital signal proessing (DSP), both in theory and hardware, we avoided the divide operator, as hardware didn't do the divide operation well. In spite of that, DSP technology advanced in both theory and hardware, finding new uses and new users. These new uses demanded high-performance math engines. Lately, new terms such as *cloud computing,* the *Internet of Things, big data, smart sensors,* etc. are driving us even harder than those initial drivers of DSP (speech, modems, hard disk drives, and three-dimensional graphics). Now, with this as a background, I can answer the question of what does a successful career look like? It is a career where the technology I helped to create became a societal necessity within the span of my career. For many of us, this has happened multiple times. It will continue to happen. All we need to do is to continue to look for those new uses and new users and then make it happen.

Author

Gene A. Frantz (Gene.Frantz@octavosystems.com) is an engineering manager/professor in practice at Rice University, Houston, Texas. He took this position after 39 years at Texas Instruments (TI), where he retired as TI's Principal Fellow. He is a recognized leader in DSP technology both within TI and throughout the industry. He holds 48 patents in the area of memories, speech, consumer products, and DSP. He has written more than 100 papers and articles and continually presents at universities and conferences worldwide. He is an IEEE Fellow.

Shan He

A successful career with signal processing training is one where you can utilize your analytical skill obtained during the training to either directly solve a technical problem, such as working as an engineer, or assist others to clarify their solution and to obtain rights associated with solution, for example, working in the patent law area. I am currently practicing patent law, and I found my signal processing background bring me tremendous advantage over other patent practitioners. This is because my strong technical background enables me to understand the invention quicker and deeper, which allows me to assist inventors to achieve the broadest possible legal protection for their invention.

Author

Shan He (shanhe@gmail.com) received her Ph.D. degree in 2007 from the University of Maryland, majoring in communications and signal processing. She worked as a research scientist in the research lab of Thomson Multimedia for three years. She then switched her career path in 2010 to become a patent agent with Lee & Hayes, PLLC, serving clients including the world's most valuable technology companies. She expects to obtain her law degree in December 2017.

Hing Cheung So

From my point of view, a successful engineer is an excellent problem solver. To solve a problem, the first step is to identify it and investigate if it is worth tackling. The second step is to formulate

the problem—describe it clearly with unambiguous requirements. Next, we apply our knowledge as well as creativity to devise solutions and then choose the best among the proposed alternatives according to the preset criteria or via balancing all the pros and cons. Finally, the solution is put into practice.

In fact, the problem-solving skill set is well trained through fundamental signal processing courses including Signals and Systems and Digital Signal Processing. For example, we learn that problems in linear time-invariant systems can be solved by either a time- or transform-domain approach, and a digital system can be designed using different finite impulse-response or infinite impulse-response filters that meet the same specifications. In the former, we also experience that the time-domain solver is easier in certain scenarios and vice versa, while one filter can be implemented with minimum complexity in the latter, stimulating us to think about the optimum choice. In addition, to be successful, I believe we should only focus on the most investable problems (i.e., think big) and realize the best solution in an efficient and persistent manner. A spirit of humility, open-mindedness, and willingness to learn is important, too.

Author

Hing Cheung So (h.c.so@cityu.edu.hk) received his Ph.D. degree in electronic engineering from The Chinese University of Hong Kong. He is a professor in the Department of Electronic Engineering, City University of Hong Kong. From 1990 to 1991, he was an electronic engineer with the Research and Development Division, Everex Systems Engineering Ltd., Hong Kong. He has been on the editorial boards of *IEEE Signal Processing Magazine, IEEE Transactions on Signal Processing, Signal Processing*, and *Digital Signal Processing*. He is a Fellow of the IEEE.

Pramod K. Varshney

Signal processing is involved in a very wide variety of systems and applications, and a person trained in this field can have a broad impact. Possibilities include hardware, software, and algorithmic developments in the areas of defense, security, health, education, quality of life, and even social good. Since signal processing training prepares one to tackle a broad range of problems, a successful career will include agility and the ability to learn quickly so as to contribute to ever-changing technological trends and needs. The key is to be able to adapt and move to new areas. When I look back at my career, with my training in statistical signal processing, I have been able to contribute to wide-ranging applications such as intelligent radars deployed on several U.S. Air Force platforms, fault detection for health management of air and space vehicles, mammography automation, and securing wireless sensor networks. In my opinion, a successful career would be one in which signal processing training is applied to solve diverse problems so as to impact societal needs and improve quality of life.

Author

Pramod K. Varshney (varshney@syr.edu) received his Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign. He is with Syracuse University, New York, where he is currently a distinguished professor of electrical engineering and computer science and the Director of the Center for Advanced Systems and Engineering. He is also an adjunct professor of radiology at Upstate Medical University, Syracuse. He received the IEEE 2012 Judith A. Resnik Award, Doctor of Engineering Honoris causa from Drexel University in 2014, and the ECE Distinguished Alumni Award from the University of Illinois in 2015. He was the president of the International Society of Information Fusion during 2001 and is a Fellow of the IEEE.

SP

New benefit from the IEEE Signal Processing Society

SPS Resource Center

The SPS Resource Center is the new home for the IEEE Signal Processing Society's online library of tutorials, lectures, presentations, and more. Unrestricted access to our fast-growing archive is now included with your SPS membership.

http://rc.signalprocessingsociety.org

We accept submissions, too!
Interested in submitting your educational materials?

sps-resourcecenter@ieee.org

Digital Object Identifier 10.1109/MSP.2017.2710838

SPECIAL REPORTS

John Edwards

Innovative Sensors Promise Longer and Healthier Lives

Signal processing leads to devices that provide faster and more insightful monitoring and diagnoses

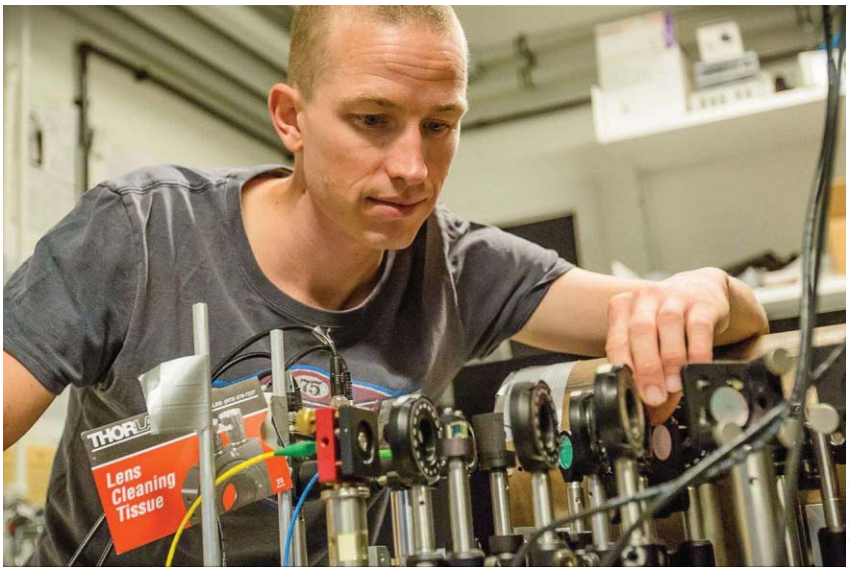
We live in a world full of sensors, and sensors are changing how we live and, more significantly, how long we live.

The health-care and medical applications sensor market is projected to expand at a compound annual growth rate of 13.1% between 2016 and 2022, according to a report issued in March 2017 by the research firm Frost & Sullivan. A key factor driving sensor sales is the growing availability of consumer and clinical devices that use sensor technology to diagnose, monitor, and track disease and fitness.

Within the next few years, an emerging generation of smaller, less expensive, and highly sophisticated sensors will find their way into a wide range of personal and professional devices. With more patient care moving out of hospitals, the use of sensor-enabled home diagnostic and monitoring devices is expected to soar, the report notes. The market for sensors used in wearable health and fitness devices is also poised to grow rapidly.

As sensor demand grows, research incorporating signal processing is leading to the development of innovative sensors designed to provide noninvasive diagnostics of different diseases, reliably monitor body functions and measure the impact of medications and activities on the human body.

Digital Object Identifier 10.1109/MSP.2017.2697158
Date of publication: 11 July 2017



NIELS BOHR INSTITUTE

FIGURE 1. Assistant Prof. Kasper Jensen investigates optical magnetic field sensor technology in a laboratory at the Niels Bohr Institute's Center for Quantum Optics.

Magnetic nerve field sensing

The human body is controlled by electrical impulses. These signals create ultraweak magnetic fields that physicians could potentially use to diagnose various diseases. Niels Bohr Institute researchers recently succeeded in developing an optical magnetic field sensor that promises to provide extremely precise measurements of weak magnetic fields emitted by nerve signals within real-world environments.

Until now, minute magnetic fields generated by nerves within a human body could only be detected with very sensitive superconducting magnetic field sensors cooled by liquid helium to near absolute zero (–273 °C). But the Niels Bohr researchers were able to create a far more practical optical magnetic field sensor that's capable of functioning at both room and body temperatures.

"We have a small glass container—1 mm × 1 mm × 8 mm—which is filled with cesium gas," says research term member Kasper Jensen, an assistant professor at the institute's Center for Quantum Optics (Figure 1). Each cesium atom rotates around itself, with the axis acting like a tiny bar magnet. When a sensor incorporating the container is held close to a nerve that's emitting an electrical pulse, it detects the magnetic field, which causes a change in the tilt

of the cesium atoms' axes. By sending a laser beam through the gas, it becomes possible to read the nerve signals' ultrasmall magnetic fields. Recent laboratory tests conducted by the researchers showed that it is possible to use the sensor to detect the magnetic field in a frog's sciatic nerve, which resembles nerves in the human body.

A magnetometer-type sensor can be used for the noninvasive diagnostics of various afflictions, including brain and heart diseases. "The key point is that a magnetometer placed outside the human body can detect signals from organs inside the human body," Jensen says. "The magnetometer does not have to touch the human body, and it is, therefore, a noninvasive method."

The sensor's operation relies on both quantum mechanics and atomic physics. "Each cesium atom has a quantum mechanical property called 'spin,' and, due to this property, the atom responds to magnetic fields," Jensen says. "One can think of the total spin of all the cesium atoms in the glass container as one large vector that points in a certain direction." When a magnetic field arrives, the spin-vector changes its direction.

"The spin-vector's direction can be measured optically with laser light," Jensen continues. "The light is detected with a photodiode, and from this detected signal we can determine the direction of the spin-vector and the magnetic field." The actual detected signal is the photo-detector's output voltage. "From that voltage we need to do some signal processing to get information about the magnetic field," Jensen says.

The photo-detector has a high bandwidth (greater than 10 MHz), and the computer-based data-acquisition card the researchers use offers a high sampling rate (fixed to 10 MHz). "To avoid aliasing, we placed a 1.9-MHz low-pass filter in between the photo-detector and the data-acquisition card," Jensen says. "The data is acquired and then processed, visualized and saved with [National Instruments] LabVIEW program."

The nerve impulse itself is relatively slow, corresponding to dc –2-kHz frequency components. "We do not really need the high sampling rate that the data-acquisition card provides," Jensen says. "Our LabVIEW program, therefore, bins the data." This action reduces the amount of data (in megabytes), enabling further data analysis to be accomplished faster. "We now have a time-signal $S(t)$, which has been low-pass filtered and binned," Jensen says. "That time-signal is saved to the computer, and we do further analysis using [The MathWorks] MATLAB software."

The researchers' experiments are run in two modes: pulsed and continuous. The data analysis is different for each mode. "In the continuous mode, we need to do a deconvolution procedure to calculate the magnetic field $B(t)$ from the time-signal $S(t)$," Jensen says. "We deconvolve with the response function: $f(t) - \cos(\Omega t)\exp(-t/T)$." The response function tells the researchers how the spin of the cesium atoms responds to a magnetic field. If there is a short pulse of magnetic field, the spin will start to oscillate at the frequency $\Omega = 400$ Hz and then decay exponentially with the time constant $T = 0.5$ ms (numbers are approximate). "In the pulsed mode, we calculate the Fourier transform of $S(t)$ and find the Fourier component at a specific frequency—in our case $\Omega = 400$ Hz," Jensen says. "The amplitude of the 400 Hz component tells us whether the nerve impulse was there or not."

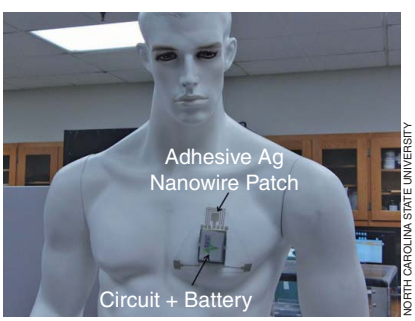
A key factor driving sensor sales is the growing availability of consumer and clinical devices that use sensor technology to diagnose, monitor, and track disease and fitness.

Jensen says the researchers did consider using a software lock-in amplifier for data analysis. "Compared to a Fourier transform, lock-in detection can be useful when one has a phase-stable signal," Jensen remarks. "We tried it out a bit but abandoned it as the phase of the signals we were looking for was changing in a way we did not fully understand." The researchers also pondered using a wavelet analysis. "This was, however, complicated by the fact that we did not know in advance the exact temporal shape of the signals we were looking for," Jensen says.

Jensen is optimistic that the technology will eventually find multiple real-world diagnostic applications. "However, so far, our main focus has been to do basic and applied research, and we do not have our magnetometers for sale," he states.

Sensing skin hydration

North Carolina State University researchers have developed a wearable, wireless sensor that can monitor a person's skin hydration to detect dehydration before it can begin posing health issues. The lightweight, flexible, and stretchable device can be built into devices that are worn on the wrist or attached to the body as a chest patch (Figure 2). "It turns out that measuring hydration of the human body is challenging, making it hard to make quantitative measurements," says research team member John Muth, a professor of electrical and computer engineering at North Carolina State University. In athletic training, for instance, the typical approach has been for an athlete measuring himself or herself, without any clothes, before and after activity. "This provides a measure of the change in hydration, since the weight change corresponds to water loss," Muth says. In clinical settings, however, a caregiver typically makes a relatively qualitative assessment simply by looking at the patient or by pulling some skin and seeing how rapidly it relaxes.



NORTH CAROLINA STATE UNIVERSITY

FIGURE 2. North Carolina State University researchers have developed a wearable, wireless sensor that can monitor a person's skin hydration to detect dehydration before it can begin posing health issues.

Health and medicine experts have long known that the skin's dielectric properties change with hydration. Existing desktop hydration measurement systems typically use a rigid probe pressed against the skin to determine impedance. Yet such calculations tend to vary in accordance to the amount of pressure applied. "Our innovations were to develop a conformal, somewhat stretchable, electrode that can be worn against the body and to miniaturize the electronics," Muth says. The new sensor includes two electrodes that are constructed out of an elastic polymer composite containing conductive silver nanowires to monitor impedance. Since the skin's electrical properties change in a predictable manner based on the individual's hydration state, the electrodes can tell how hydrated the skin is. The entire system is about the size of an Apple Watch.

Muth estimates that adding the sensor to a wearable smart device would cost only about US$1.

The device offers multiple potential applications, Muth notes. "High-performance athletes would like to know more about their hydration state when training, since this can be directly linked to athletic performance," he explains. "First responders can dehydrate when working in extreme conditions." In tests performed on custom-made artificial skins incorporating a wide range of hydration levels, the researchers found that the wearable sensor's performance was unaffected by ambient humidity.

An Analog Devices 5933 network analyzer chip handles most of the signal processing. "When the skin is in contact with the electrode, we are looking for a change in impedance that is connected to the hydration state of the skin," he says. "The network analyzer chip approach allows us to measure the complex impedance as a function of frequency."

The chip uses direct digital syntheses to produce a sinusoidal output voltage at a known frequency and amplitude that is then applied to the electrode. "The voltage across the electrode is received and amplified, and passes through a low pass filter," Muth says. "A discrete Fourier transform (DFT) is performed for each frequency in the sweep, storing both the real and imaginary components of the DFT result." The impedance is then calculated by multiplying a scaling factor obtained by measuring a known impedance by one over the magnitude of the DFT result. "The phase angle in radians is calculated by taking the arctangent of the ratio of the imaginary and real parts," Muth says. "Once the magnitude of the impedance and phase angle are known, the resistive and reactive components are calculated for each frequency."

Once an individual measurement has been made, a variety of techniques can be used to average the data or to detect specific events, such as the onset of sweating. "We still need to investigate how motion artifacts can influence the data," Muth says. "Knowledge of how to fuse other data, such as the body temperature, external humidity, heart rate or other parameters, could also be useful since often people are also interested in these other parameters."

Both the watch and patch can wirelessly transmit sensor data to external devices, allowing data to be monitored by the user or a designated third party, such as a doctor in a hospital or clinic. Muth estimates that adding the sensor to a wearable smart device would cost only about US$1.

UNIVERSITY OF TEXAS AT DALLAS

FIGURE 3. Shalini Prasad (right), professor of bioengineering at the University of Texas at Dallas, and doctoral student Rujuta Munje have designed a wearable, flexible biosensor that can reliably detect and quantify glucose from very small amounts of human perspiration. A close-up of the sensor is shown in the top-left corner.

Monitoring glucose via perspiration

Can a person's glucose level be quickly and conveniently monitored through skin perspiration? That was the question University of Texas at Dallas researchers sought to answer as they began designing a wearable device that could be used by individuals with diabetes, or at risk of developing the disease, to measure their blood sugar levels.

Shalini Prasad, a University of Texas at Dallas professor of bioengineering, and doctoral student Rujuta Munje recently demonstrated a sensor they designed to reliably detect and quantify glucose in human perspiration (Figure 3). Conventional patient-type blood glucose readers use a small blood sample, typically obtained via a finger prick. The new textile-based sensor, however, detects glucose from a tiny amount of

ambient perspiration on a person's skin. "Our sensor mechanism uses the same chemistry and enzymatic reaction found in blood glucose testing strips," Prasad says. "Our design, however, accounts for the low volume of ambient sweat typically present in areas such as under a wrist device or patch."

The new device requires perspiration volumes of under a microliter—approximately equal the amount of liquid that would fit into a cube the size of a salt crystal—to make an accurate measurement that's then displayed on a digital readout, according to Prasad. The sensor is based on an off-the-shelf polymer-based textile material. The current prototype is a small, flexible, rod-shaped device measuring about an inch long. "The innovation is that we positioned the electrodes onto the textile in a manner that allows a very small volume of sweat to spread effectively through the surface," Prasad says.

The researchers turned to Kalman filtering to differentiate readings. "The Kalman filter is one that works very well for dynamic systems that have a lot of uncertainty associated with them," Prasad remarks. "You apply Kalman filtering to a particular sector to try to establish, with a great degree of certainty, whether glucose is the molecule that is specifically interacting with the sensor surface or whether the current change that's happening is due to something else."

According to Prasad, sensor calibration response was calculated using $n = 4$ samples. The response to the varying glucose concentration was captured in terms of percentage change in total impedance (Z_{mod}) between the baseline step impedance and the impedance obtained for that particular concentration. The Z_{mod} was captured at 100 Hz, the highest signal over noise ratio. Specific signal threshold (SST) was estimated by measuring replicates of a blank buffer sample and calculating the mean result and standard deviation. The noise level was defined as the three times of standard deviation in baseline (zero dose) measurement. Limit of detection was identified as the lowest glucose concentration likely to be reliably distinguished from the SST and at which detection is feasible. "We have shown that this particular sensor works robustly not just in a lab environment, but kind of in a translation environment as well," Prasad says. It can adjust itself to variations in environmental condition such as temperature, humidity, the people who are wearing it, and so forth."

The researchers are now looking toward refining the sensor into a device that could potentially replace blood sample-based glucose readers. "We believe it could easily be incorporated into existing consumer electronics platforms," Prasad says.

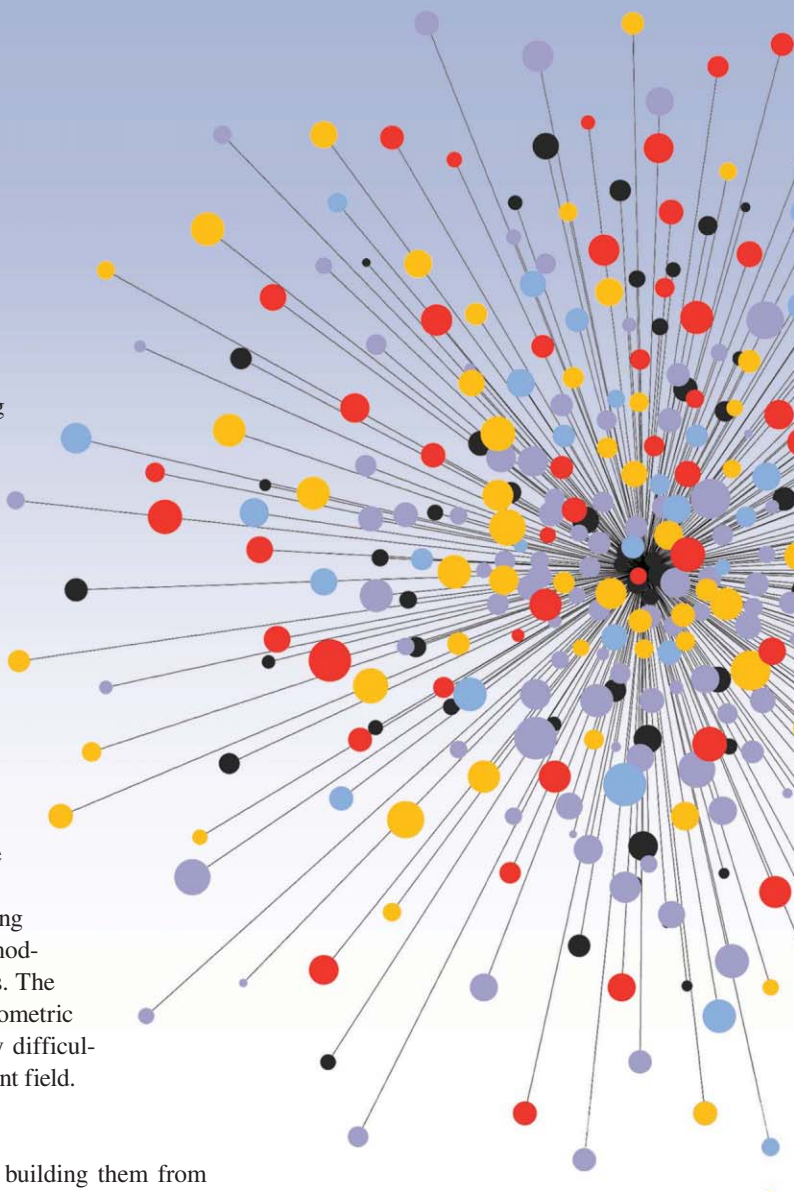
Author

John Edwards (jedwards@johnedwardsmedia.com) is a technolgy writer based in the Phoenix, Arizona, area.

SP

Digital Object Identifier 10.1109/MSP.2017.2721488

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst



Many scientific fields study data with an underlying structure that is non-Euclidean. Some examples include social networks in computational social sciences, sensor networks in communications, functional networks in brain imaging, regulatory networks in genetics, and meshed surfaces in computer graphics. In many applications, such geometric data are large and complex (in the case of social networks, on the scale of billions) and are natural targets for machine-learning techniques. In particular, we would like to use deep neural networks, which have recently proven to be powerful tools for a broad range of problems from computer vision, natural-language processing, and audio analysis. However, these tools have been most successful on data with an underlying Euclidean or grid-like structure and in cases where the invariances of these structures are built into networks used to model them.

Geometric deep learning is an umbrella term for emerging techniques attempting to generalize (structured) deep neural models to non-Euclidean domains, such as graphs and manifolds. The purpose of this article is to overview different examples of geometric deep-learning problems and present available solutions, key difficulties, applications, and future research directions in this nascent field.

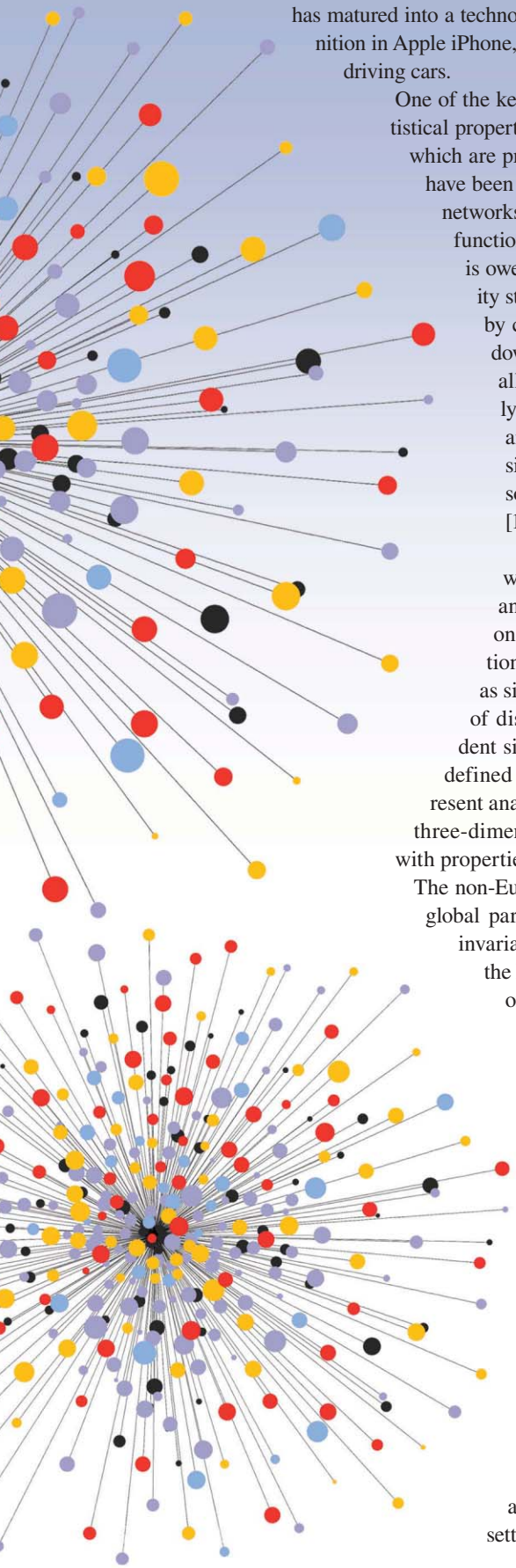
Overview of deep learning

Deep learning refers to learning complicated concepts by building them from simpler ones in a hierarchical or multilayer manner. Artificial neural networks are popular realizations of such deep multilayer hierarchies. In the past few years, the growing computational power of modern graphics processing unit (GPU)-based computers and the availability of large training data sets have allowed successfully training neural networks with many layers and degrees of freedom (DoF) [1]. This has led to qualitative breakthroughs on a wide variety of tasks, from speech recognition [2], [3] and machine translation [4] to image analysis and computer vision [5]–[11] (see [12]

Geometric Deep Learning

Going beyond Euclidean data

Digital Object Identifier 10.1109/MSP.2017.2693418
Date of publication: 11 July 2017



and [13] for many additional examples of successful applications of deep learning). Today, deep learning has matured into a technology that is widely used in commercial applications, including Siri speech recognition in Apple iPhone, Google text translation, and Mobileye vision-based technology for autonomously driving cars.

One of the key reasons for the success of deep neural networks is their ability to leverage statistical properties of the data, such as stationarity and compositionality through local statistics, which are present in natural images, video, and speech [14], [15]. These statistical properties have been related to physics [16] and formalized in specific classes of convolutional neural networks (CNNs) [17]–[19]. In image analysis applications, one can consider images as functions on the Euclidean space (plane), sampled on a grid. In this setting, stationarity is owed to shift invariance, locality is due to the local connectivity, and compositionality stems from the multiresolution structure of the grid. These properties are exploited by convolutional architectures [20], which are built of alternating convolutional and downsampling (pooling) layers. The use of convolutions has a twofold effect. First, it allows extracting local features that are shared across the image domain and greatly reduces the number of parameters in the network with respect to generic deep architectures (and thus also the risk of overfitting), without sacrificing the expressive capacity of the network. Second, the convolutional architecture itself imposes some priors about the data, which appear very suitable especially for natural images [17]–[19], [21].

While deep-learning models have been particularly successful when dealing with speech, image, and video signals, in which there are an underlying Euclidean structure, recently there has been a growing interest in trying to apply learning on non-Euclidean geometric data. Such kinds of data arise in numerous applications. For instance, in social networks, the characteristics of users can be modeled as signals on the vertices of the social graph [22]. Sensor networks are graph models of distributed interconnected sensors, whose readings are modeled as time-dependent signals on the vertices. In genetics, gene expression data are modeled as signals defined on the regulatory network [23]. In neuroscience, graph models are used to represent anatomical and functional structures of the brain. In computer graphics and vision, three-dimensional (3-D) objects are modeled as Riemannian manifolds (surfaces) endowed with properties such as color texture.

The non-Euclidean nature of such data implies that there are no such familiar properties as global parameterization, common system of coordinates, vector space structure, or shift invariance. Consequently, basic operations like convolution that are taken for granted in the Euclidean case are even not well defined on non-Euclidean domains. The purpose of this article is to show different methods of translating the key ingredients of successful deep-learning methods, such as CNNs, to non-Euclidean data.

Geometric learning problems

Broadly speaking, we can distinguish between two classes of geometric learning problems. In the first class of problems, the goal is to characterize the structure of the data. The second class of problems deals with analyzing functions defined on a given non-Euclidean domain. These two classes are related, because understanding the properties of functions defined on a domain conveys certain information about the domain, and vice versa, the structure of the domain imposes certain properties on the functions on it.

Structure of the domain

As an example of the first class of problems, assume to be given a set of data points with some underlying low-dimensional structure embedded into a high-dimensional Euclidean space. Recovering that low-dimensional structure is often referred to as *manifold learning* or *nonlinear dimensionality reduction* and is an instance of unsupervised learning (note that the notion of manifold in this setting can be considerably more general than a classical smooth manifold; see, e.g.,

[24] and [25]). Many methods for nonlinear dimensionality reduction consist of two steps: first, they start with constructing a representation of local affinity of the data points (typically, a sparsely connected graph). Second, the data points are embedded into a low-dimensional space, trying to preserve some criterion of the original affinity. For example, spectral embeddings tend to map points with many connections between them to nearby locations, and multidimensional scaling (MDS)-type methods try to preserve global information, such as graph geodesic distances. Examples of manifold learning include different flavors of MDS [26], locally linear embedding [27], stochastic neighbor embedding [28], spectral embeddings, such as Laplacian eigenmaps [29] and diffusion maps [30], and deep models [31]. Instead of embedding the vertices, the graph structure can be processed by decomposing it into small subgraphs called *motifs* [36] or *graphlets* [37]. Finally, most recent approaches [32]–[34] tried to apply the successful word-embedding model [35] to graphs.

Today, deep learning has matured into a technology that is widely used in commercial applications.

In some cases, the data are presented as a manifold or graph at the outset, and the first step of constructing the affinity structure described previously is unnecessary. For instance, in computer graphics and vision applications, one can analyze 3-D shapes represented as meshes by constructing local geometric descriptors capturing, e.g., curvature-like properties [38], [39]. In social network analysis applications the topological structure of the social graph representing the social relations between people carries important insights allowing, e.g., to classify the vertices and detect communities [40]. In natural-language processing, words in a corpus can be represented by the co-occurrence graph, where two words are connected if they often appear near each other [41].

Data on a domain

Our second class of problems deals with analyzing functions defined on a given non-Euclidean domain. We can further break down such problems into two subclasses: problems where the domain is fixed and those where multiple domains are given. For example, assume that we are given the geographic coordinates of the users of a social network, represented as a time-dependent signal on the vertices of the social graph. An important application in location-based social networks is to predict the position of the user given his or her past behavior as well as that of his or her friends [42]. In this problem, the domain (social graph) is assumed to be fixed; methods of signal processing on graphs, which have previously been reviewed in *IEEE Signal Processing Magazine* [43], can be applied to this setting, in particular, to define an operation similar to convolution in the spectral domain. This, in turn, allows generalizing CNN models to graphs [44], [45]. In computer graphics and vision applications, finding similarity and correspondence between shapes are examples of the second subclass of problems: each shape is modeled as a manifold, and one has to work with multiple such domains. In this setting, a generalization of convolution in the spatial domain using local charting [46]–[48] appears to be more appropriate.

Brief history

The main focus of this review is on this second class of problems, namely, learning functions on non-Euclidean structured domains, and, in particular, attempts to generalize the popular CNNs to such settings. The first attempts to generalize neural networks to graphs we are aware of are due to Gori et al. [49], who proposed a scheme combining recurrent neural networks (RNNs) and random walk models. This approach went almost unnoticed, reemerging in a modern form in [50] and [51] due to the renewed recent interest in deep learning. The first formulation of CNNs on graphs is due to Bruna et al. [52], who used the definition of convolutions in the spectral domain. Their article, while being of conceptual importance, came with significant computational drawbacks that fell short of a truly useful method. These drawbacks were subsequently addressed in the follow-up works of Henaff et al. [44] and Defferrard et al. [45]. In the latter article, graph CNNs (GCNNs) allowed achieving some state-of-the-art results.

In a parallel effort in the computer vision and graphics community, Masci et al. [47] showed the first CNN model on meshed surfaces, resorting to a spatial definition of the convolution operation based on local intrinsic patches. Among other applications, such models were shown to achieve state-of-the-art performance in finding correspondence between deformable 3-D shapes. Follow-up works proposed different construction of intrinsic patches on point clouds [48], [53] and general graphs [54].

The interest in deep learning on graphs or manifolds has exploded in the past year, resulting in numerous attempts to apply these methods to a broad spectrum of problems ranging from biochemistry [55] to recommender systems [56]. Because such applications originate in different fields that usually do not cross-fertilize, publications in this domain tend to use different terminology and notation, making it difficult for a newcomer to grasp the foundations and current state-of-the-art methods. We believe that our article comes at the right time, attempting to systemize and bring some order into the field.

Signal processing, differential geometry, and graph theory

Geometric deep-learning frameworks dealt with in this paper are based on notions in differential geometry and graph theory. Unfortunately, these topics are insufficiently known in the signal processing community, and to our knowledge, there is no introductory-level reference treating these so different structures in a common way. One of our goals is to provide an accessible overview of these models, resorting as much as possible to the intuition of traditional signal processing.

One of the key differences between Euclidean and non-Euclidean learning settings is the lack of traditional operations such as convolutions. Various non-Euclidean convolutional architectures differ in the way a convolution-like operation is

formulated on graphs and manifolds. One way is to resort to the analogy of the convolution theorem, defining the convolution in the spectral domain. An alternative is to think of the convolution as a template matching in the spatial domain. Such a distinction is, however, far from being clear-cut: as we will see, some approaches draw their formulation from the spectral domain, essentially boiling down to applying filters in the spatial domain. It is also possible to combine these two approaches, resorting to spatio-frequency analysis techniques, such as wavelets or the windowed Fourier transform. We have provided sidebars to illustrate important concepts, and Table 1 lists the notations used throughout the article. Additional materials, data, and examples of code are available at geometricdeeplearning.com. Table 2 provides a summary of the geometric deep-learning methods presented in this article.

Table 1. The notations used in this article.

Notation	
$\mathbb{R}^m$	m-dimensional Euclidean space
$a, \mathbf{a}, \mathbf{A}$	Scalar, vector, matrix
$\bar{a}$	Complex conjugate of a
Ω, x	Arbitrary domain, coordinate on it
$f \in L^2(\Omega)$	Square-integrable function on Ω
$\delta_{x'}(x), \delta_{ij}$	Delta function at x', Kronecker delta
$\{f_i, y_i\}_{i \in I}$	Training set
$\mathcal{T}_v$	Translation operator
$\tau, \mathcal{L}_\tau$	Deformation field, operator
$\hat{f}$	Fourier transform of f
$f \star g$	Convolution of f and g
$\mathcal{X}, T\mathcal{X}, T_x\mathcal{X}$	Manifold, its tangent bundle, tangent space at x
$\langle \cdot, \cdot, \rangle_{T\mathcal{X}}$	Riemannian metric
$f \in L^2(\mathcal{X})$	Scalar field on manifold $\mathcal{X}$
$F \in L^2(T\mathcal{X})$	Tangent vector field on manifold $\mathcal{X}$
A^*	Adjoint of operator A
∇, div, Δ	Gradient, divergence, Laplace operators
$\mathcal{V}, \mathcal{E}, \mathcal{F}$	Vertices and edges of a graph, faces of a mesh
$w_{ij}, \mathbf{W}$	Weight matrix of a graph
$f \in L^2(\mathcal{V})$	Functions on vertices of a graph
$F \in L^2(\mathcal{E})$	Functions on edges of a graph
ϕ_i, λ_i	Laplacian eigenfunctions, eigenvalues
$h_t(\cdot, \cdot)$	Heat kernel
$\mathbf{\Phi}_k$	Matrix of first k Laplacian eigenvectors
$\mathbf{\Lambda}_k$	Diagonal matrix of first k Laplacian eigenvalues
ξ	Pointwise nonlinearity (ReLU)
$\gamma_{l,l'}(x), \Gamma_{l,l'}$	Convolutional filter in spatial and spectral domain

Deep learning on Euclidean domains

Geometric priors

Consider a compact d-dimensional Euclidean domain $\Omega = [0,1]^d \subset \mathbb{R}^d$ on which square-integrable functions $f \in L^2(\Omega)$ are defined (e.g., in image analysis applications, images can be thought of as functions on the unit square $\Omega = [0,1]^2$). We consider a generic supervised learning setting, in which an unknown function $y : L^2(\Omega) \to \mathcal{Y}$ is observed on a training set

$$\{f_i \in L^2(\Omega), y_i = y(f_i)\}_{i \in I}. \tag{1}$$

In a supervised classification setting, the target space $\mathcal{Y}$ can be thought discrete, with $|\mathcal{Y}|$ being the number of classes. In a multiple object recognition setting, we can replace $\mathcal{Y}$ by a multi-K-dimensional simplex, which represents the posterior class probabilities $p(y|x)$. In regression tasks, we may consider $\mathcal{Y} = \mathbb{R}^m$. In the vast majority of computer-vision and speech-analysis tasks, there are several crucial prior assumptions on the unknown function y. As we will see in the following sections, these assumptions are effectively exploited by CNN architectures.

Stationarity

Let

$$\mathcal{T}_v f(x) = f(x - v), \qquad x, v \in \Omega, \tag{2}$$

be a translation operator acting on functions $f \in L^2(\Omega)$ [we assume periodic boundary conditions to ensure that the operation is well defined over $L^2(\Omega)$]. Our first assumption is that the function y is either invariant or equivariant with respect to translations, depending on the task. In the former case, we have $y(\mathcal{T}_v f) = y(f)$ for any $f \in L^2(\Omega)$ and $v \in \Omega$. This is typically the case in object classification tasks. In the latter, we have $y(\mathcal{T}_v f) = \mathcal{T}_v y(f)$, which is well defined when the output of the model is a space in which translations can act (e.g., in problems of object localization, semantic segmentation, or motion estimation). Our definition of invariance should not be confused with the traditional notion of translation invariant systems in signal processing, which corresponds to translation equivariance in our language (because the output translates whenever the input translates).

Local deformations and scale separation

Similarly, a deformation $\mathcal{L}_\tau$, where $\tau : \Omega \to \Omega$ is a smooth vector field, acts on $L^2(\Omega)$ as $\mathcal{L}_\tau f(x) = f(x - \tau(x))$. Deformations can model local translations, changes in point of view, rotations, and frequency transpositions [18]. Most tasks studied in computer vision are not only translation invariant/equivariant but also stable with respect to local deformations [57], [18]. In tasks that are translation invariant, we have

$$|y(\mathcal{L}_\tau f) - y(f)| \approx \|\nabla \tau\|, \tag{3}$$

CNN Architecture

CNNs are currently among the most successful deep-learning architectures in a variety of tasks; in particular, in computer vision. A typical CNN used in computer-vision applications (see Figure S1) consists of multiple convolutional layers (6), passing the input image through a set of filters Γ followed by pointwise nonlinearity ξ (typically, half-rectifiers $\xi(z) = \max(0, z)$ are used, although practitioners have experimented with a diverse range of choices [13]). The model can also include a bias term, which is equivalent to adding a constant coordinate to the input.

A network composed of K convolutional layers put together $U(f) = (C_{\Gamma^{(K)}} \ldots \circ C_{\Gamma^{(2)}} \circ C_{\Gamma^{(1)}})(f)$ produces pixel-wise features that are covariant with respect to translation and approximately covariant to local deformations. Typical computer-vision applications requiring covariance are semantic image segmentation [8] or motion estimation [59].

In applications requiring invariance, such as image classification [7], the convolutional layers are typically interleaved with pooling layers (8) progressively reducing the resolution of the image passing through the network. Alternatively, one can integrate the convolution and downsampling in a single linear operator (convolution with stride). Recently, some authors have also experimented with convolutional layers that increase the spatial resolution using interpolation kernels [60]. These kernels can be learned efficiently by mimicking the so-called *algorithme à trous* [61], also referred to as *dilated convolution*.

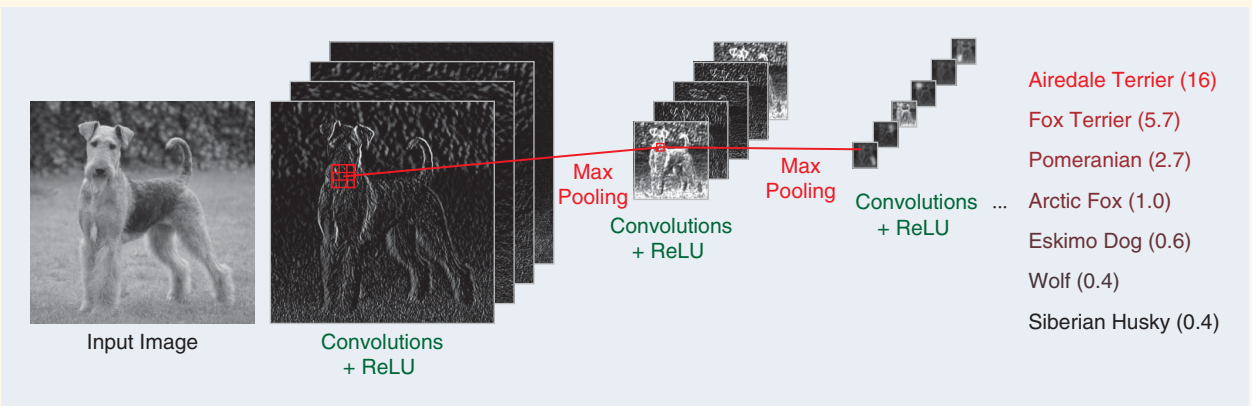


FIGURE S1. The typical CNN architecture used in computer-vision applications such as image classification.

Table 2. The dichotomy of geometric deep-learning methods.

Method	Type	Data
SCNN [52]	Spectral	Graph
GCNN/ChebNet [45]	Spectrum free	Graph
GCN [77]	Spectrum free	Graph
GNN [78]	Spectrum free	Graph
Geodesic CNN [47]	Charting	Mesh
Anisotropic CNN [48]	Charting	Mesh/point cloud
MoNet [54]	Charting	Graph/mesh/point cloud
Localized SCNN [89]	Combined	Mesh/point cloud

for all f, τ. Here, $\|\nabla\tau\|$ measures the smoothness of a given deformation field. In other words, the quantity to be predicted does not change much if the input image is slightly deformed. In tasks that are translation equivariant, we have

$$|y(\mathcal{L}_\tau f) - \mathcal{L}_\tau y(f)| \approx \|\nabla\tau\|. \quad (4)$$

This property is much stronger than the previous one, because the space of local deformations has a high dimensionality, as opposed to the d-dimensional translation group. It follows from (3) that we can extract sufficient statistics at a lower spatial resolution by downsampling demodulated localized filter responses without losing approximation power. An important consequence of this is that long-range dependencies can be broken into multiscale local interaction terms, leading to hierarchical models in which spatial resolution is progressively reduced. To illustrate this principle, denote by

$$Y(x_1, x_2; v) = \text{Prob}(f(u) = x_1 \text{ and } f(u+v) = x_2) \quad (5)$$

the joint distribution of two image pixels at an offset v from each other. In the presence of long-range dependencies, this joint distribution will not be separable for any v. However, the deformation stability prior states that $Y(x_1, x_2; v) \approx Y(x_1, x_2; v(1+\epsilon))$ for small ϵ. In other words,

whereas long-range dependencies indeed exist in natural images and are critical to object recognition, they can be captured and downsampled at different scales. This principle of stability to local deformations has been exploited in the computer-vision community in models other than CNNs, for instance, deformable parts models [58]. In practice, the Euclidean domain Ω is discretized using a regular grid with n points; the translation and deformation operators are still well defined so the above properties also hold in the discrete setting.

CNNs

Stationarity and stability to local translations are both leveraged in CNNs (see "CNN Architecture" and [1], [12], [13], and references therein for a more in-depth review of CNNs and their applications.) A CNN consists of several convolutional layers of the form $\mathbf{g} = C_\Gamma(\mathbf{f})$, acting on a p-dimensional input $\mathbf{f}(x) = (f_1(x), \ldots, f_p(x))$ by applying a bank of filters $\Gamma = (\gamma_{l,l'}), l = 1, \ldots, q, l' = 1, \ldots, p$ and pointwise nonlinearity ξ,

$$g_l(x) = \xi\left(\sum_{l'=1}^{p} (f_{l'} \star \gamma_{l,l'})(x)\right), \tag{6}$$

producing a q-dimensional output $\mathbf{g}(x) = (g_1(x), \ldots, g_q(x))$ often referred to as the *feature maps*. Here,

$$(f \star \gamma)(x) = \int_\Omega f(x - x')\gamma(x')\,dx' \tag{7}$$

denotes the standard convolution. According to the local deformation prior, the filters Γ have compact spatial support.

Additionally, a downsampling or pooling layer $\mathbf{g} = P(\mathbf{f})$ may be used, defined as

$$g_l(x) = P(\{f_l(x'): x' \in \mathcal{N}(x)\}), \quad l = 1, \ldots, q, \tag{8}$$

where $\mathcal{N}(x) \subset \Omega$ is a neighborhood around x and P is a permutation-invariant function, such as an L_p-norm (in the latter case, the choice of $p = 1$, 2, or ∞ results in average, energy, or max pooling).

A convolutional network is constructed by composing several convolutional and optionally pooling layers, obtaining a generic hierarchical representation

$$U_\Theta(f) = (C_{\Gamma^{(K)}} \cdots P \cdots \circ C_{\Gamma^{(2)}} \circ C_{\Gamma^{(1)}})(f), \tag{9}$$

where $\Theta = \{\Gamma^{(1)}, \ldots, \Gamma^{(K)}\}$ is the hypervector of the network parameters (all the filter coefficients). The model is said to be deep if it comprises multiple layers, though this notion is rather vague, and one can find examples of CNNs with as few as a couple and as many as hundreds of layers [11]. The output features enjoy translation invariance/covariance depending on whether spatial resolution is progressively lost by means of pooling or kept fixed. Moreover, if one specifies the convolutional tensors to be complex wavelet decomposition operators and uses complex modulus as pointwise nonlinearities, one can provably obtain stability to local deformations [17]. Although this stability is not rigorously proved for generic compactly supported convolutional tensors, it underpins the empirical success of CNN architectures across a variety of computer-vision applications [1].

In supervised learning tasks, one can obtain the CNN parameters by minimizing a task-specific cost L on the training set $\{f_i, y_i\}_{i \in I}$,

$$\min_\Theta \sum_{i \in I} L(U_\Theta(f_i), y_i), \tag{10}$$

for instance, $L(x, y) = \|x - y\|$. If the model is sufficiently complex and the training set is sufficiently representative, when applying the learned model to previously unseen data, one expects $U(f) \approx y(f)$. Although (10) is a nonconvex optimization problem, stochastic optimization methods offer excellent empirical performance. Understanding the structure of the optimization problems (10) and finding efficient strategies for its solution is an active area of research in deep learning [62]–[66].

A key advantage of CNNs explaining their success in numerous tasks is that the geometric priors on which CNNs are based result in a learning complexity that avoids the curse of dimensionality. Thanks to the stationarity and local deformation priors, the linear operators at each layer have a constant number of parameters, independent of the input size n (number of pixels in an image). Moreover, thanks to the multiscale hierarchical property, the number of layers grows at a rate $O(\log n)$, resulting in a total learning complexity of $O(\log n)$ parameters.

The geometry of manifolds and graphs

Our main goal is to generalize CNN-type constructions to non-Euclidean domains. In this article, by non-Euclidean domains, we refer to two prototypical structures: manifolds and graphs. While arising in very different fields of mathematics (differential geometry and graph theory, respectively), in our context, these structures share several common characteristics that we will try to emphasize throughout our review.

Manifolds

Roughly, a manifold is a space that is locally Euclidean. One of the simplest examples is a spherical surface modeling our planet: around a point, it seems to be planar, which has led generations of people to believe in the flatness of the Earth. Formally speaking, a (differentiable) d-dimensional manifold $\mathcal{X}$ is a topological space where each point x has a neighborhood that is topologically equivalent (homeomorphic) to a d-dimensional Euclidean space, called the *tangent space* and denoted by $T_x\mathcal{X}$ [see Figure 1(a)]. The collection of tangent spaces at all points (more formally, their disjoint union) is referred to as the *tangent bundle* and denoted by $T\mathcal{X}$. On each tangent space, we define an inner product $\langle \cdot, \cdot \rangle_{T_x\mathcal{X}}: T_x\mathcal{X} \times T_x\mathcal{X} \to \mathbb{R}$, which is additionally assumed to depend smoothly on the position x. This inner product is

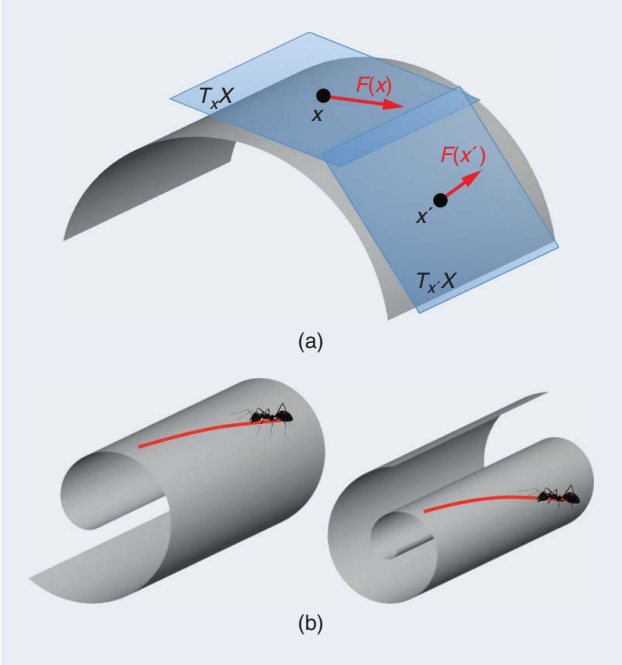


FIGURE 1. (a) The tangent space and tangent vectors on a 2-D manifold (surface). (b) Examples of isometric deformations.

called a *Riemannian metric* in differential geometry and allows performing local measurements of angles, distances, and volumes. A manifold equipped with a metric is called a *Riemannian manifold*.

It is important to note that the definition of a Riemannian manifold is completely abstract and does not require a geometric realization in any space. However, a Riemannian manifold can be realized as a subset of a Euclidean space (in which case it is said to be embedded in that space) by using the structure of the Euclidean space to induce a Riemannian metric. The celebrated Nash embedding theorem guarantees that any sufficiently smooth Riemannian manifold can be realized in a Euclidean space of sufficiently high dimension [67]. An embedding is not necessarily unique; two different realizations of a Riemannian metric are called *isometries*.

Two-dimensional (2-D) manifolds (surfaces) embedded into $\mathbb{R}^3$ are used in computer graphics and vision to describe boundary surfaces of 3-D objects, colloquially referred to as *3-D shapes*. This term is somewhat misleading because *3-D* here refers to the dimensionality of the embedding space rather than that of the manifold. Thinking of such a shape as made of infinitely thin material, inelastic deformations that do not stretch or tear it are isometric. Isometries do not affect the metric structure of the manifold, and consequently, they preserve any quantities that can be expressed in terms of the Riemannian metric (called *intrinsic*). Conversely, properties pertaining to the specific realization of the manifold in the Euclidean space are called *extrinsic*. As an intuitive illustration of this difference, imagine an insect that lives on a 2-D surface [Figure 1(b)]. The surface can be placed in the Euclidean space in any way, but as long as it is transformed isometrically, the insect would not notice any difference. The insect in fact does not even know of the existence of the embedding space, as its only world is 2-D. This is an intrinsic viewpoint. A human observer, on the other hand, sees a surface in 3-D space—this is an extrinsic point of view.

Calculus on manifolds

Our next step is to consider functions defined on manifolds. We are particularly interested in two types of functions: A scalar field is a smooth real function $f:\mathcal{X}\to\mathbb{R}$ on the manifold. A tangent vector field $F:\mathcal{X}\to T\mathcal{X}$ is a mapping attaching a tangent vector $F(x)\in T_x\mathcal{X}$ to each point x. As we will see in the following, tangent vector fields are used to formalize the notion of infinitesimal displacements on the manifold. We define the Hilbert spaces of scalar and vector fields on manifolds, denoted by $L^2(\mathcal{X})$ and $L^2(T\mathcal{X})$, respectively, with the following inner products:

$$\langle f,g\rangle_{L^2(\mathcal{X})}=\int_{\mathcal{X}} f(x)\,g(x)\,dx, \tag{11}$$

$$\langle F,G\rangle_{L^2(T\mathcal{X})}=\int_{\mathcal{X}}\langle F(x),G(x)\rangle_{T_x\mathcal{X}}\,dx. \tag{12}$$

Here, dx denotes a d-dimensional volume element induced by the Riemannian metric.

In calculus, the notion of derivative describes how the value of a function changes with an infinitesimal change of its argument. One of the big differences distinguishing classical calculus from differential geometry is a lack of vector space structure on the manifold, prohibiting us from naively using expressions like $f(x+dx)$. The conceptual leap that is required to generalize such notions to manifolds is the need to work locally in the tangent space.

To this end, we define the differential of f as an operator $df:T\mathcal{X}\to\mathbb{R}$ acting on tangent vector fields. At each point x, the differential can be defined as a linear functional $df(x)=\langle\nabla f(x),\cdot\rangle_{T_x\mathcal{X}}$ acting on tangent vectors $F(x)\in T_x\mathcal{X}$, which model a small displacement around x. The change of the function value as the result of this displacement is given by applying the functional to the tangent vector, $df(x)\,F(x)=\langle\nabla f(x),F(x)\rangle_{T_x\mathcal{X}}$, and can be thought of as an extension of the notion of the classical directional derivative.

The operator $\nabla f:L^2(\mathcal{X})\to L^2(T\mathcal{X})$ in the previous definition is called the *intrinsic gradient* and is similar to the classical notion of the gradient defining the direction of the steepest change of the function at a point, with the only difference that the direction is now a tangent vector. Similarly, the intrinsic divergence is an operator $\mathrm{div}:L^2(T\mathcal{X})\to L^2(\mathcal{X})$ acting on tangent vector fields and is (formal) adjoint to the gradient operator [71],

$$\langle F,\nabla f\rangle_{L^2(T\mathcal{X})}=\langle\nabla^* F,f\rangle_{L^2(\mathcal{X})}=\langle-\mathrm{div}F,f\rangle_{L^2(\mathcal{X})}. \tag{13}$$

Physically, a tangent vector field can be thought of as a flow of material on a manifold. The divergence measures the net flow of a field at a point, allowing to distinguish between field sources and sinks. Finally, the Laplacian (or Laplace–Beltrami

Physical Interpretation of Laplacian Eigenfunctions

Given a function f on the domain $\mathcal{X}$, the Dirichlet energy

$$\varepsilon_{\text{Dir}}(f) = \int_{\mathcal{X}} \|\nabla f(x)\|^2_{T_x\mathcal{X}}\, dx = \int_{\mathcal{X}} f(x)\,\Delta f(x)\, dx, \tag{S1}$$

measures how smooth it is [the last identity in (S1) stems from (15)]. We are looking for an orthonormal basis on $\mathcal{X}$, containing k smoothest possible functions (Figure S2), by solving the optimization problem

$$\begin{aligned} &\min_{\phi_0} \mathcal{E}_{\text{Dir}}(\phi_0) \ \text{ s.t. } \ \|\phi_0\| = 1 \\ &\min_{\phi_i} \mathcal{E}_{\text{Dir}}(\phi_i) \ \text{ s.t. } \ \|\phi_i\| = 1, i = 1, 2, \ldots k-1 \\ &\qquad\qquad \phi_i \perp \text{span}\{\phi_0, \ \ldots, \phi_{i-1}\}. \end{aligned} \tag{S2}$$

In the discrete setting, when the domain is sampled at n points, (S2) can be rewritten as

$$\min_{\mathbf{\Phi}_k \in \mathbb{R}^{n\times k}} \text{trace}(\mathbf{\Phi}_k^\top \Delta \mathbf{\Phi}_k) \ \text{ s.t. } \ \mathbf{\Phi}_k^\top \mathbf{\Phi}_k = \mathbf{I}, \tag{S3}$$

where $\mathbf{\Phi}_k = (\boldsymbol{\phi}_0, \ldots \boldsymbol{\phi}_{k-1})$. The solution of (S3) is given by the first k eigenvectors of Δ satisfying

$$\Delta \mathbf{\Phi}_k = \mathbf{\Phi}_k \mathbf{\Lambda}_k, \tag{S4}$$

where $\mathbf{\Lambda}_k = \text{diag}(\lambda_0, \ldots, \lambda_{k-1})$ is the diagonal matrix of corresponding eigenvalues. The eigenvalues $0 = \lambda_0 \le \lambda_1 \le \cdots \lambda_{k-1}$ are nonnegative due to the positive semidefiniteness of the Laplacian and can be interpreted as frequencies, where $\phi_0 = \text{const}$ with the corresponding eigenvalue $\lambda_0 = 0$ plays the role of the direct current component.

The Laplacian eigendecomposition can be carried out in two ways. First, (S4) can be rewritten as a generalized eigenproblem $(\mathbf{D} - \mathbf{W})\mathbf{\Phi}_k = \mathbf{A}\mathbf{\Phi}_k\mathbf{\Lambda}_k$, resulting in $\mathbf{A}$-orthogonal eigenvectors, $\mathbf{\Phi}_k^\top \mathbf{A} \mathbf{\Phi}_k = \mathbf{I}$. Alternatively, introducing a change of variables $\mathbf{\Psi}_k = \mathbf{A}^{1/2}\mathbf{\Phi}_k$, we can obtain a standard eigendecomposition problem $\mathbf{A}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{A}^{-1/2}\mathbf{\Psi}_k = \mathbf{\Psi}_k\mathbf{\Lambda}_k$ with orthogonal eigenvectors $\mathbf{\Psi}_k^\top \mathbf{\Psi}_k = \mathbf{I}$. When $\mathbf{A} = \mathbf{D}$ is used, the matrix $\Delta = \mathbf{A}^{-1/2}(\mathbf{D} - \mathbf{W})\mathbf{A}^{-1/2}$ is referred to as the *normalized symmetric Laplacian*.

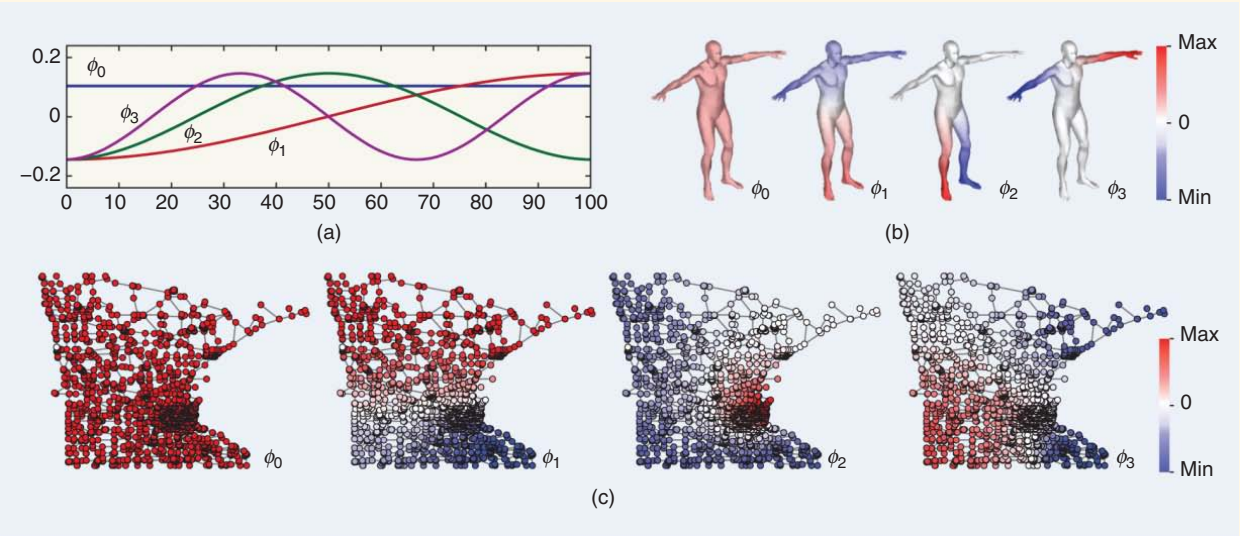


FIGURE S2. An example of the first four Laplacian eigenfunctions $\phi_0, \ldots, \phi_3$ on (a) a Euclidean domain (1-D line), and (b) and (c) non-Euclidean domains [(b) a human shape modeled as a 2-D manifold, and (c) a Minnesota road graph]. In the Euclidean case, the result is the standard Fourier basis comprising sinusoids of increasing frequency. In all cases, the eigenfunction ϕ_0 corresponding to zero eigenvalue is constant (direct current component).1-D: one-dimensional.

operator in differential geometric jargon) $\Delta: L^2(\mathcal{X}) \to L^2(\mathcal{X})$ is an operator,

$$\Delta f = -\text{div}(\nabla f), \tag{14}$$

acting on scalar fields. Employing relation (13), it is easy to see that the Laplacian is self-adjoint (symmetric),

$$\langle \nabla f, \nabla f \rangle_{L^2(T\mathcal{X})} = \langle \Delta f, f \rangle_{L^2(\mathcal{X})} = \langle f, \Delta f \rangle_{L^2(\mathcal{X})}. \tag{15}$$

The left-hand-side in (15) is known as the Dirichlet energy in physics and measures the smoothness of a scalar field on the manifold (see "Physical Interpretation of Laplacian Eigenfunctions"). The Laplacian can be interpreted as the difference between the average of a function on an infinitesimal

Heat Diffusion on Non-Euclidean Domains

An important application of spectral analysis and, historically, the main motivation for its development by Joseph Fourier, is the solution of partial differential equations. Here, we are particularly interested in heat propagation on non-Euclidean domains. This process is governed by the heat diffusion equation, which in the simplest setting of homogeneous and isotropic diffusion has the form

$$\begin{cases} f_t(x,t) = -c\Delta f(x,t) \\ f(x,0) = f_0(x) \ \text{(Initial condition)} \end{cases} \tag{S5}$$

with additional boundary conditions if the domain has a boundary. $f(x,t)$ represents the temperature at point x at time t. Equation (S5) encodes Newton's law of cooling, according to which the rate of temperature change of a body (left-hand side) is proportional to the difference between its own temperature and that of the surrounding right-hand side. The proportion coefficient c is referred to as the *thermal diffusivity constant*; here, we assume it to be equal to one for the sake of simplicity. The solution of (S5) is given by applying the heat operator $H^t = e^{-t\Delta}$ to the initial condition and can be expressed in the spectral domain as

$$\begin{aligned} f(x,t) = e^{-t\Delta} f_0(x) &= \sum_{i\geq 0} \langle f_0, \phi_i \rangle_{L^2(\mathcal{X})} e^{-t\lambda_i} \phi_i(x) \\ &= \int_{\mathcal{X}} f_0(x') \underbrace{\sum_{i\geq 0} e^{-t\lambda_i} \phi_i(x) \phi_i(x')}_{h_t(x,x')} dx'. \end{aligned} \tag{S6}$$

$h_t(x,x')$ is known as the *heat kernel* (Figure S3) and represents the solution of the heat equation with an initial condition $f_0(x) = \delta_{x'}(x)$, or, in signal processing terms, an *impulse response*. In physical terms, $h_t(x,x')$ describes how much heat flows from a point x to point x' in time t. In the Euclidean case, the heat kernel is shift invariant, $h_t(x,x') = h_t(x-x')$, allowing to interpret the integral in (S6) as a convolution $f(x,t) = (f_0 \star h_t)(x)$. In the spectral domain, convolution with the heat kernel amounts to low-pass filtering with frequency response $e^{-t\lambda}$. Larger values of diffusion time t result in lower effective cutoff frequency and thus smoother solutions in space (corresponding to the intuition that longer diffusion smoothes more the initial heat distribution).

The crosstalk between two heat kernels positioned at points x and x' allows to measure an intrinsic distance

$$d_t^2(x,x') = \int_{\mathcal{X}} (h_t(x,y) - h_t(x',y))^2 dy \tag{S7}$$

$$= \sum_{i\geq 0} e^{-2t\lambda_i} (\phi_i(x) - \phi_i(x'))^2 \tag{S8}$$

referred to as the *diffusion distance* [30]. Note that when interpreting (S7) and (S8) as spatial- and frequency-domain norms $\|\cdot\|_{L^2(\mathcal{X})}$ and $\|\cdot\|_{\ell^2}$, respectively, their equivalence is the consequence of the Parseval identity. Unlike geodesic distance that measures the length of the shortest path on the manifold or graph, the diffusion distance has an effect of averaging over different paths. It is thus more robust to perturbations of the domain, e.g., introduction or removal of edges in a graph or cuts on a manifold.

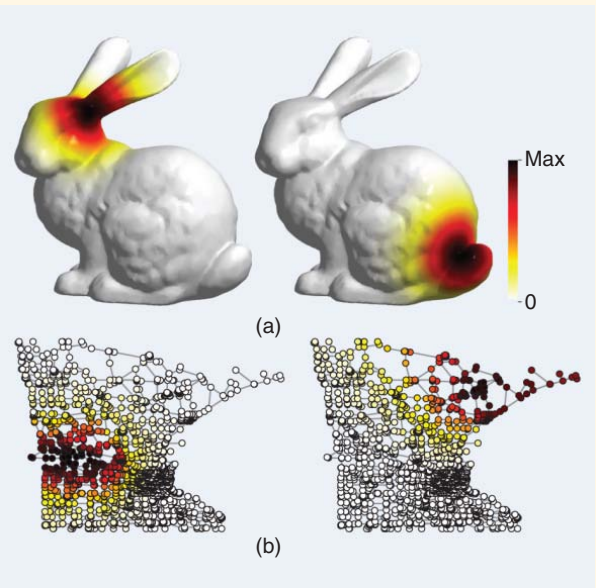


FIGURE S3. The examples of heat kernels on non-Euclidean domains [(a) manifold, and (b) graph]. Observe how moving the heat kernel to a different location changes its shape, which is an indication of the lack of shift invariance.

sphere around a point and the value of the function at the point itself. It is one of the most important operators in mathematical physics, used to describe phenomena as diverse as heat diffusion (see "Heat Diffusion on Non-Euclidean Domains"), quantum mechanics, and wave propagation. As we will see in the following, the Laplacian plays a central role in signal processing and learning on non-Euclidean domains, as its eigenfunctions generalize the classical Fourier bases, allowing to perform spectral analysis on manifolds and graphs.

It is important to note that all the previous definitions are coordinate free. By defining a basis in the tangent space, it is possible to express tangent vectors as d-dimensional vectors and the Riemannian metric as a $d \times d$ symmetric positive-definite matrix.

Graphs and discrete differential operators

Another type of constructions we are interested in are graphs, which are popular models of networks, interactions, and

similarities between different objects. For simplicity, we will consider weighted undirected graphs, formally defined as a pair $(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{1,\ldots,n\}$ is the set of n vertices, and $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$ is the set of edges, where the graph being undirected implies that $(i,j)\in\mathcal{E}$ if $(j,i)\in\mathcal{E}$. Furthermore, we associate a weight $a_i>0$ with each vertex $i\in\mathcal{V}$, and a weight $w_{ij}\geq 0$ with each edge $(i,j)\in\mathcal{E}$.

Real functions $f:\mathcal{V}\to\mathbb{R}$ and $F:\mathcal{E}\to\mathbb{R}$ on the vertices and edges of the graph, respectively, are roughly the discrete analogy of continuous scalar and tangent vector fields in differential geometry (it is tacitly assumed here that F is alternating, i.e., $F_{ij}=-F_{ji}$). We define Hilbert spaces $L^2(\mathcal{V})$ and $L^2(\mathcal{E})$ of such functions by specifying the respective inner products,

$$\langle f,g\rangle_{L^2(\mathcal{V})}=\sum_{i\in\mathcal{V}}a_i f_i g_i, \tag{16}$$

$$\langle F,G\rangle_{L^2(\mathcal{E})}=\sum_{i\in\mathcal{E}}w_{ij}F_{ij}G_{ij}. \tag{17}$$

Let $f\in L^2(\mathcal{V})$ and $F\in L^2(\mathcal{E})$ be functions on the vertices and edges of the graphs, respectively. We can define differential operators acting on such functions analogously to differential operators on manifolds [72]. The graph gradient is an operator $\nabla:L^2(\mathcal{V})\to L^2(\mathcal{E})$ mapping functions defined on vertices to functions defined on edges,

$$(\nabla f)_{ij}=f_i-f_j, \tag{18}$$

automatically satisfying $(\nabla f)_{ij}=-(\nabla f)_{ji}$. The graph divergence is an operator $\mathrm{div}:L^2(\mathcal{E})\to L^2(\mathcal{V})$ doing the converse,

$$(\mathrm{div}F)_i=\frac{1}{a_i}\sum_{j:(i,j)\in\mathcal{E}}w_{ij}F_{ij}. \tag{19}$$

It is easy to verify that the two operators are adjoint with respect to the inner products (16) and (17),

$$\langle F,\nabla f\rangle_{L^2(\mathcal{E})}=\langle\nabla^* F,f\rangle_{L^2(\mathcal{V})}=\langle-\mathrm{div}F,f\rangle_{L^2(\mathcal{V})}. \tag{20}$$

The graph Laplacian is an operator $\Delta:L^2(\mathcal{V})\to L^2(\mathcal{V})$ defined as $\Delta=-\mathrm{div}\nabla$. Combining definitions (18) and (19), it can be expressed in the familiar form

$$(\Delta f)_i=\frac{1}{a_i}\sum_{(i,j)\in\mathcal{E}}w_{ij}(f_i-f_j). \tag{21}$$

Note that (21) captures the intuitive geometric interpretation of the Laplacian as the difference between the local average of a function around a point and the value of the function at the point itself.

Denoting by $\mathbf{W}=(w_{ij})$ the $n\times n$ matrix of edge weights [it is assumed that $w_{ij}=0$ if $(i,j)\notin\mathcal{E}$], by $\mathbf{A}=\mathrm{diag}(a_1,\ldots,a_n)$ the diagonal matrix of vertex weights, and by $\mathbf{D}=\mathrm{diag}\left(\sum_{j:j\neq i}w_{ij}\right)$ the degree matrix, the graph Laplacian application to a function $f\in L^2(\mathcal{V})$ represented as a column vector $\mathbf{f}=(f_1,\ldots,f_n)^\top$ can be written in matrix-vector form as

$$\Delta\mathbf{f}=\mathbf{A}^{-1}(\mathbf{D}-\mathbf{W})\mathbf{f}. \tag{22}$$

The choice of $\mathbf{A}=\mathbf{I}$ in (22) is referred to as the *unnormalized graph Laplacian*; another popular choice is $\mathbf{A}=\mathbf{D}$ producing the random walk Laplacian [73].

Discrete manifolds

As previously mentioned, there are many practical situations in which one is given a sampling of points arising from a manifold but not the manifold itself. In computer graphics applications, reconstructing a correct discretization of a manifold from a point cloud is a difficult problem of its own, referred to as *meshing* (see "Laplacian on Discrete Manifolds"). In manifold-learning problems, the manifold is typically approximated as a graph capturing the local affinity structure. We stress that the term *manifold* as used in the context of generic data science is not geometrically rigorous and can have less structure than a classical smooth manifold we have defined beforehand. For example, a set of points that looks locally Euclidean in practice may have self-intersections, infinite curvature, different dimensions depending on the scale and location at which one looks, extreme variations in density, and noise with confounding structure.

Fourier analysis on non-Euclidean domains

The Laplacian operator is a self-adjoint positive-semidefinite operator, admitting on a compact domain an eigendecomposition with a discrete set of orthonormal eigenfunctions $\phi_0,\phi_1,\ldots$ (satisfying $\langle\phi_i,\phi_j\rangle_{L^2(\mathcal{X})}=\delta_{ij}$) and nonnegative real eigenvalues $0=\lambda_0\leq\lambda_1\leq\ldots$ (referred to as the *spectrum* of the Laplacian),

$$\Delta\phi_i=\lambda_i\phi_i, i=0,1,\ldots. \tag{23}$$

[Note that in the Euclidean case, the Fourier transform of a function defined on a finite interval (which is a compact set) or its periodic extension is discrete. In practical settings, all domains we are dealing with are compact.]

The eigenfunctions are the smoothest functions in the sense of the Dirichlet energy (see "Physical Interpretation of Laplacian Eigenfunctions") and can be interpreted as a generalization of the standard Fourier basis [given, in fact, by the eigenfunctions of the one-dimensional (1-D) Euclidean Laplacian, $-\left(d^2/x^2\right)e^{i\omega x}=\omega^2 e^{i\omega x}$] to a non-Euclidean domain. It is important to emphasize that the Laplacian eigenbasis is intrinsic due to the intrinsic construction of the Laplacian itself.

A square-integrable function f on $\mathcal{X}$ can be decomposed into Fourier series as

$$f(x)=\sum_{i\geq 0}\underbrace{\langle f,\phi_i\rangle_{L^2(\mathcal{X})}}_{\hat{f}_i}\phi_i(x), \tag{24}$$

where the projection on the basis functions producing a discrete set of Fourier coefficients $(\hat{f}_0,\hat{f}_1,\ldots)$ generalizes the analysis (forward transform) stage in classical signal processing,

Laplacian on Discrete Manifolds

In computer graphics and vision applications, 2-D manifolds are commonly used to model 3-D shapes. There are several common ways of discretizing such manifolds. First, the manifold is assumed to be sampled at n points. Their embedding coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are referred to as a *point cloud*. Second, a graph is constructed upon these points, acting as its vertices. The edges of the graph represent the local connectivity of the manifold, telling whether two points belong to a neighborhood or not. The graph can be endowed, e.g., with Gaussian-edge weights

$$w_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2}. \tag{S9}$$

This simplest discretization, however, does not correctly capture the geometry of the underlying continuous manifold (e.g., the graph Laplacian would typically not converge to the continuous Laplacian operator of the manifold with the increase of the sampling density [68]). A geometrically consistent discretization is possible with an additional structure of faces $\mathcal{F} \in \mathcal{V} \times \mathcal{V} \times \mathcal{V}$, where $(i,j,k) \in \mathcal{F}$ implies $(i,j),(i,k),(k,j) \in \mathcal{E}$. The collection of faces represents the underlying continuous manifold as a polyhedral surface consisting of small triangles glued together. The triplet $(\mathcal{V}, \mathcal{E}, \mathcal{F})$ is referred to as *triangular mesh*. To be a correct discretization of a manifold (a manifold mesh), every edge must be shared by exactly two triangular faces; if the manifold has a boundary, any boundary edge must belong to exactly one triangle.

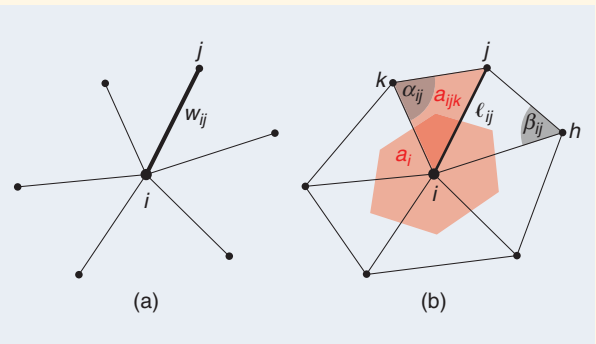


FIGURE S4. The two commonly used discretizations of a 2-D manifold: (a) an undirected graph and (b) a triangular mesh.

On a triangular mesh, the simplest discretization of the Riemannian metric is given by assigning each edge a length $\ell_{ij} > 0$, which must additionally satisfy the triangle inequality in every triangular face. The mesh Laplacian is given by (21) with

$$w_{ij} = \frac{-\ell_{ij}^2 + \ell_{jk}^2 + \ell_{ik}^2}{8a_{ijk}} + \frac{-\ell_{ij}^2 + \ell_{jh}^2 + \ell_{ih}^2}{8a_{ijh}}; \tag{S10}$$

$$a_i = \frac{1}{3} \sum_{jk:(i,j,k) \in \mathcal{F}} a_{ijk}, \tag{S11}$$

where $a_{ijk} = \sqrt{s_{ijk}(s_{ijk} - \ell_{ij})(s_{ijk} - \ell_{jk})(s_{ijk} - \ell_{ik})}$ is the area of triangle ijk given by the Heron formula, and $s_{ijk} = (1/2)(\ell_{ij} + \ell_{jk} + \ell_{ki})$ is the semiperimeter of triangle ijk. The vertex weight a_i is interpreted as the local area element (shown in red in Figure S4). Note that the weights (S10) and (S11) are expressed solely in terms of the discrete metric ℓ and are thus intrinsic. When the mesh is infinitely refined under some technical conditions, such a construction can be shown to converge to the continuous Laplacian of the underlying manifold [69].

An embedding of the mesh (amounting to specifying the vertex coordinates $\mathbf{x}_1, \ldots, \mathbf{x}_n$) induces a discrete metric $\ell_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$, whereby (S10) become the cotangent weights

$$w_{ij} = \frac{1}{2}(\cot\alpha_{ij} + \cot\beta_{ij}) \tag{S12}$$

ubiquitously used in computer graphics [70].

and summing up the basis functions with these coefficients is the synthesis (inverse transform) stage.

A centerpiece of classical Euclidean signal processing is the property of the Fourier transform diagonalizing the convolution operator, colloquially referred to as the *convolution theorem*. This property allows to express the convolution $f \star g$ of two functions in the spectral domain as the elementwise product of their Fourier transforms,

$$\widehat{(f \star g)}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx \int_{-\infty}^{\infty} g(x) e^{-i\omega x} dx. \tag{25}$$

Unfortunately, in the non-Euclidean case, we cannot even define the operation $x - x'$ on the manifold or graph, so the notion of convolution (7) does not directly extend to this case. One possibility to generalize convolution to non-Euclidean domains is by using the convolution theorem as a definition,

$$(f \star g)(x) = \sum_{i \geq 0} \langle f, \phi_i \rangle_{L^2(X)} \langle g, \phi_i \rangle_{L^2(X)} \phi_i(x). \tag{26}$$

One of the key differences of such a construction from the classical convolution is the lack of shift invariance. In terms of signal processing, it can be interpreted as a position-dependent filter. While parameterized by a fixed number of coefficients in the frequency domain, the spatial representation of the filter can vary dramatically at different points (see Figure S3).

In situations where the graph is constructed from the data, a straightforward choice of the edge weights (S9) of the graph is the covariance of the data Let $\mathbf{F}$ denote the input data distribution and

$$\mathbf{\Sigma} = \mathbb{E}(\mathbf{F} - \mathbb{E}\mathbf{F})(\mathbf{F} - \mathbb{E}\mathbf{F})^\top \tag{S13}$$

be the data covariance matrix. If each point has the same variance $\sigma_{ii} = \sigma^2$, then diagonal operators on the Laplacian simply scale the principal components of $\mathbf{F}$.

In natural images, because their distribution is approximately stationary, the covariance matrix has a circulant structure $\sigma_{ij} \approx \sigma_{i-j}$ and is thus diagonalized by the standard discrete cosine transform (DCT) basis. It follows that the principal components of $\mathbf{F}$ roughly correspond to the DCT basis vectors ordered by frequency. Moreover, natural images exhibit a power spectrum $\mathbb{E}|\hat{f}(\omega)|^2 \sim |\omega|^{-2}$, because nearby pixels are more correlated than faraway pixels [14]. It results that principal components of the covariance are essentially ordered from low to high frequencies, which is consistent with the standard group structure of the Fourier basis. When applied to natural images represented as graphs with weights defined by the covariance, the SCNN construction recovers the standard CNN, without any prior knowledge [76] (Figure S5). Indeed, the linear operators $\mathbf{\Phi\Gamma}_{l,l'}\mathbf{\Phi}^\top$ in (27) are by the previous argument diagonal in the Fourier basis, hence translation invariant, hence classical convolutions. Furthermore, the "Spectrum-Free Methods" section explains how spatial subsampling can also be obtained via dropping the last part of the spectrum of the Laplacian, leading to pooling, and ultimately to standard CNNs.

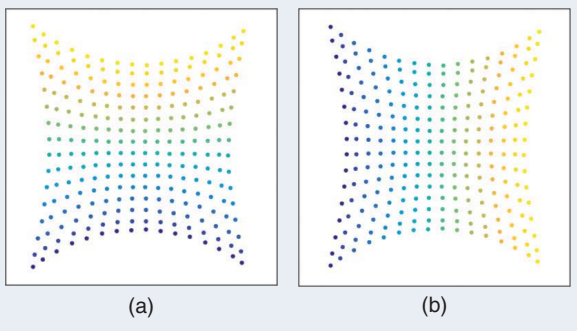


FIGURE S5. The 2-D embedding of pixels in 16 × 16 image patches using a Euclidean radial basis function (RBF) kernel. The RBF kernel is constructed as in (S9), by using the covariance σ_{ij} as Euclidean distance between two features. The pixels are embedded in a 2-D space using the first two eigenvectors of the resulting graph Laplacian. The colors in (a) and (b) represent the horizontal and vertical coordinates of the pixels, respectively. The spatial arrangement of pixels is roughly recovered from correlation measurements.

The previous discussion also applies to graphs instead of manifolds, where one only has to replace the inner product in (24) and (26) with the discrete one (16). All of the sums over i would become finite, as the graph Laplacian matrix Δ has n eigenvectors. In matrix-vector notation, the generalized convolution $f \star g$ can be expressed as $\mathbf{Gf} = \mathbf{\Phi}\,\mathrm{diag}(\hat{\mathbf{g}})\mathbf{\Phi}^\top\mathbf{f}$, where $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_n)$ is the spectral representation of the filter, and $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)$ denotes the Laplacian eigenvectors (S8). The lack of shift invariance results in the absence of circulant (Toeplitz) structure in the matrix $\mathbf{G}$, which characterizes the Euclidean setting. Furthermore, it is easy to see that the convolution operation commutes with the Laplacian, $\mathbf{G\Delta f} = \mathbf{\Delta Gf}$.

Uniqueness and stability

Finally, it is important to note that the Laplacian eigenfunctions are not uniquely defined. To start with, they are defined up to sign, i.e., $\Delta(\pm\phi) = \lambda(\pm\phi)$. Thus, even isometric domains might have different Laplacian eigenfunctions. Furthermore, if a Laplacian eigenvalue has multiplicity, then the associated eigenfunctions can be defined as orthonormal basis spanning the corresponding eigensubspace (or said differently, the eigenfunctions are defined up to an orthogonal transformation in the

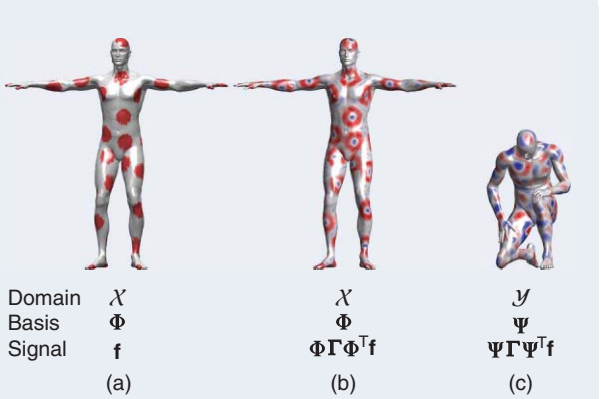


FIGURE 2. A toy example illustrating the difficulty of generalizing spectral filtering across non-Euclidean domains. (a) A function defined on a manifold (function values are represented by color). (b) The result of the application of an edge-detection filter in the frequency domain. (c) The same filter applied on the same function but on a different (nearly isometric) domain produces a completely different result. The reason for this behavior is that the Fourier basis is domain dependent and the filter coefficients learned on one domain cannot be applied to another one in a straightforward manner.

Citation Network Analysis Application

The CORA citation network [90] is a graph containing 2,708 vertices representing articles and 5,429 edges representing citations (Figure S6). Each article is described by a 1,433-dimensional bag-of-words feature vector and belongs to seven classes. For simplicity, the network is treated as an undirected graph. Applying the SCNN with two spectral convolutional layers parameterized according to (37), the authors of [77] obtained classification accuracy of 81.6% (compared to the previous best result of 75.7%). In [54], this result was slightly improved further, reaching 81.7% accuracy with the use of MoNet architecture.

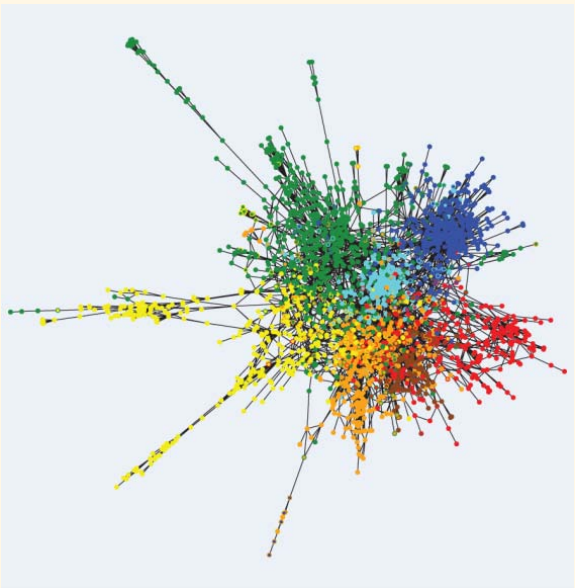


FIGURE S6. The classifying of a research article in the CORA data set with MoNet. Shown is the citation graph, where each node is an article and an edge represents a citation. Vertex fill and outline colors represent the predicted and ground-truth labels, respectively; ideally, the two colors should coincide. (Figure reproduced from [54].)

eigensubspace). A small perturbation of the domain can lead to very large changes in the Laplacian eigenvectors, especially those associated with high frequencies. At the same time, the definition of heat kernels (S6) and diffusion distances (S8) does not suffer from these ambiguities, e.g., the sign ambiguity disappears as the eigenfunctions are squared. Heat kernels also appear to be robust to domain perturbations.

Spectral methods

We have now finally gotten to our main goal, namely, constructing a generalization of the CNN architecture on non-Euclidean domains. We will start with the assumption that the domain on which we are working is fixed, and for the rest of this section, we will use the problem of classification of a signal on a fixed graph as the prototypical application. We have seen that convolutions are linear operators that commute with the Laplacian operator. Therefore, given a weighted graph, a first route to generalize a convolutional architecture is by first restricting our interest to linear operators that commute with the graph Laplacian. This property, in turn, implies operating on the spectrum of the graph weights, given by the eigenvectors of the graph Laplacian.

Spectral CNN

Similarly to the convolutional layer (6) of a classical Euclidean CNN, Bruna et al. [52] define a spectral convolutional layer as

$$\mathbf{g}_l = \xi\left(\sum_{l'=1}^{q} \mathbf{\Phi}_k \mathbf{\Gamma}_{l,l'} \mathbf{\Phi}_k^\top \mathbf{f}_{l'}\right), \tag{27}$$

where the $n \times p$ and $n \times q$ matrices $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_p)$ and $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_q)$ represent the p- and q-dimensional input and output signals on the vertices of the graph, respectively (we use $n = |\mathcal{V}|$ to denote the number of vertices in the graph), $\mathbf{\Gamma}_{l,l'}$ is a $k \times k$ diagonal matrix of spectral multipliers representing a filter in the frequency domain, and ξ is a nonlinearity applied on the vertex-wise function values. Using only the first k eigenvectors in (27) sets a cutoff frequency that depends on the intrinsic regularity of the graph and also the sample size. Typically, $k \ll n$, because only the first Laplacian eigenvectors describing the smooth structure of the graph are useful in practice.

If the graph has an underlying group invariance, such a construction can discover it. In particular, standard CNNs can be redefined from the spectral domain (see "Rediscovering Standard CNNs Using Correlation Kernels"). However, in many cases the graph does not have a group structure, or the group structure does not commute with the Laplacian, and so we cannot think of each filter as passing a template across $\mathcal{V}$ and recording the correlation of the template with that location.

We should stress that a fundamental limitation of the spectral construction is its restriction to a single domain. The reason is that spectral filter coefficients (27) are basis dependent. It implies that if we learn a filter with respect to basis $\mathbf{\Phi}_k$ on one domain, and then try to apply it on another domain with another basis $\mathbf{\Psi}_k$, the result could be very different (see Figure 2). It is possible to construct compatible orthogonal bases across different domains resorting to a joint diagonalization procedure [74], [75]. However, such a construction requires the knowledge of some correspondence between the domains. In applications like social network analysis, e.g., where dealing with two time instances of a social graph in which new vertices and edges have been added, such a correspondence can be easily computed and is therefore a reasonable assumption. Conversely, in computer graphics applications, finding correspondence between shapes is in itself a very hard problem, so assuming known correspondence between the domains is a rather unreasonable assumption.

Assuming that $k = O(n)$ eigenvectors of the Laplacian are kept, a convolutional layer (27) requires $pqk = O(n)$ parameters to train. We will see next how the global and local regularity of the graph can be combined to produce layers with constant number of parameters (i.e., such that the number of learnable parameters per layer does not depend upon the size of the input), which is the case in classical Euclidean CNNs.

The non-Euclidean analogy of pooling is graph coarsening, in which only a fraction $\alpha < 1$ of the graph vertices is retained. The eigenvectors of graph Laplacians at two different resolutions are related by the following multigrid property: let $\mathbf{\Phi}, \tilde{\mathbf{\Phi}}$ denote the $n \times n$ and $\alpha n \times \alpha n$ matrices of Laplacian eigenvectors of the original and the coarsened graph, respectively. Then,

$$\tilde{\mathbf{\Phi}} \approx \mathbf{P}\mathbf{\Phi}\begin{pmatrix}\mathbf{I}_{\alpha n}\\ 0\end{pmatrix}, \tag{28}$$

where $\mathbf{P}$ is an $\alpha n \times n$ binary matrix whose ith row encodes the position of the ith vertex of the coarse graph on the original graph. It follows that strided convolutions can be generalized using the spectral construction by keeping only the low-frequency components of the spectrum. This property also allows us to interpret (via interpolation) the local filters at deeper layers in the spatial construction to be low frequency. However, because in (27) the nonlinearity is applied in the spatial domain, in practice one has to recompute the graph Laplacian eigenvectors at each resolution and apply them directly after each pooling step.

The spectral construction (27) assigns a DoF for each eigenvector of the graph Laplacian. In most graphs, individual high-frequency eigenvectors become highly unstable. However, similarly as the wavelet construction in Euclidean domains, by appropriately grouping high-frequency eigenvectors in each octave, one can recover meaningful and stable information. As shown next, this principle also entails better learning complexity.

Spectral CNN with smooth spectral multipliers

To reduce the risk of overfitting, it is important to adapt the learning complexity to reduce the number of free parameters of the model [44], [52]. On Euclidean domains, this is achieved by learning convolutional kernels with small spatial support, which enables the model to learn a number of parameters independent of the input size. To achieve a similar learning complexity in the spectral domain, it is thus necessary to restrict the class of spectral multipliers to those corresponding to localized filters.

For that purpose, we have to express spatial localization of filters in the frequency domain. In the Euclidean case, smoothness in the frequency domain corresponds to spatial decay, because

$$\int_{-\infty}^{+\infty} |x|^{2k} |f(x)|^2 dx = \int_{-\infty}^{+\infty} \left| \frac{\partial^k \hat{f}(\omega)}{\partial \omega^k} \right|^2 d\omega, \tag{29}$$

by virtue of the Parseval identity. This suggests that, to learn a layer in which features will be not only shared across locations but also well localized in the spatial domain, one can learn spectral multipliers that are smooth. Smoothness can be prescribed by learning only a subsampled set of spectral multipliers and using an interpolation kernel to obtain the rest, such as cubic splines.

However, the notion of smoothness also requires some geometry in the spectral domain. In the Euclidean setting, such a geometry naturally arises from the notion of frequency, e.g., in the plane, the similarity between two Fourier atoms $e^{i\boldsymbol{\omega}^\top \mathbf{x}}$ and $e^{i\boldsymbol{\omega}'^\top \mathbf{x}}$ can be quantified by the distance $\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|$, where $\mathbf{x}$ denotes the 2-D planar coordinates, and $\boldsymbol{\omega}$ is the 2-D frequency vector. On graphs, such a relation can be defined by means of a dual graph with weights $\tilde{w}_{ij}$ encoding the similarity between two eigenvectors ϕ_i and ϕ_j.

A particularly simple choice consists in choosing a 1-D arrangement, obtained by ordering the eigenvectors according to their eigenvalues. [In the mentioned 2-D example, this would correspond to ordering the Fourier basis function according to the sum of the corresponding frequencies $\omega_1 + \omega_2$. Although numerical results on simple low-dimensional graphs show that the 1-D arrangement given by the spectrum of the Laplacian is efficient at creating spatially localized filters [52], an open fundamental question is how to define a dual graph on the eigenvectors of the Laplacian in which smoothness (obtained by applying the diffusion operator) corresponds to localization in the original graph.] In this setting, the spectral multipliers are parameterized as

$$\mathrm{diag}(\mathbf{\Gamma}_{l,l'}) = \mathbf{B}\boldsymbol{\alpha}_{l,l'}, \tag{30}$$

where $\mathbf{B} = (b_{ij}) = (\beta_j(\lambda_i))$ is a $k \times q$ fixed interpolation kernel [e.g., $\beta_j(\lambda)$ can be cubic splines], and $\boldsymbol{\alpha}$ is a vector of q interpolation coefficients. To obtain filters with constant spatial support (i.e., independent of the input size n), one should choose a sampling step $\gamma \sim n$ in the spectral domain, which results in a constant number $n\gamma^{-1} = O(1)$ of coefficients $\boldsymbol{\alpha}_{l,l'}$ per filter. Therefore, by combining spectral layers with graph coarsening, this model has $O(\log n)$ total trainable parameters for inputs of size n, thus recovering the same learning complexity as CNNs on Euclidean grids.

Even with such a parameterization of the filters, the spectral CNN (27) entails a high computational complexity of performing forward and backward passes, because they require an expensive step of matrix multiplication by $\mathbf{\Phi}_k$ and $\mathbf{\Phi}_k^\top$. While on Euclidean domains such a multiplication can be efficiently carried in $O(n \log n)$ operations using fast-Fourier-transform-type algorithms, for general graphs such algorithms do not exist and the complexity is $O(n^2)$. We will see next how to alleviate this cost by avoiding explicit computation of the Laplacian eigenvectors.

Spectrum-free methods

A polynomial of the Laplacian acts as a polynomial on its eigenvalues. Thus, instead of explicitly operating in the frequency

domain with spectral multipliers as in (30), it is possible to represent the filters via a polynomial expansion:

$$g_\alpha(\Delta) = \mathbf{\Phi} g_\alpha(\mathbf{\Lambda})\mathbf{\Phi}^\top, \tag{31}$$

where

$$g_\alpha(\lambda) = \sum_{j=0}^{r-1} \alpha_j \lambda^j, \tag{32}$$

$\boldsymbol{\alpha}$ is the r-dimensional vector of polynomial coefficients, and $g_\alpha(\mathbf{\Lambda}) = \mathrm{diag}(g_\alpha(\lambda_1), \ldots, g_\alpha(\lambda_n))$, resulting in filter matrices $\mathbf{\Gamma}_{l,l'} = g_{\alpha_{l,l'}}(\mathbf{\Lambda})$ whose entries have an explicit form in terms of the eigenvalues.

An important property of this representation is that it automatically yields localized filters, for the following reason. Because the Laplacian is a local operator (working on one-hop neighborhoods), the action of its jth power is constrained to j hops. Because the filter is a linear combination of powers of the Laplacian, overall (32) behaves like a diffusion operator limited to r hops around each vertex.

GCNN, also known as ChebNet

Defferrard et al. used the Chebyshev polynomials generated by the recurrence relation [45]

$$\begin{aligned} T_j(\lambda) &= 2\lambda T_{j-1}(\lambda) - T_{j-2}(\lambda), \\ T_0(\lambda) &= 1, \\ T_1(\lambda) &= \lambda. \end{aligned} \tag{33}$$

A filter (32) can thus be parameterized uniquely via an expansion of order $r-1$ such that

$$\begin{aligned} g_\alpha(\tilde{\Delta}) &= \sum_{j=0}^{r-1} \alpha_j \mathbf{\Phi} T_j(\tilde{\mathbf{\Lambda}})\mathbf{\Phi}^\top \\ &= \sum_{j=0}^{r-1} \alpha_j T_j(\tilde{\Delta}), \end{aligned} \tag{34}$$

where $\tilde{\Delta} = 2\lambda_n^{-1}\Delta - \mathbf{I}$ and $\tilde{\mathbf{\Lambda}} = 2\lambda_n^{-1}\mathbf{\Lambda} - \mathbf{I}$ denotes a rescaling of the Laplacian mapping its eigenvalues from the interval $[0, \lambda_n]$ to $[-1, 1]$ (necessary because the Chebyshev polynomials form an orthonormal basis in $[-1, 1]$).

Denoting $\bar{\mathbf{f}}^{(j)} = T_j(\tilde{\Delta})\mathbf{f}$, we can use the recurrence relation (33) to compute $\bar{\mathbf{f}}^{(j)} = 2\tilde{\Delta}\bar{\mathbf{f}}^{(j-1)} - \bar{\mathbf{f}}^{(j-2)}$, with $\bar{\mathbf{f}}^{(0)} = \mathbf{f}$ and $\bar{\mathbf{f}}^{(1)} = \tilde{\Delta}\mathbf{f}$. The computational complexity of this procedure is therefore $O(rn)$ operations and does not require an explicit computation of the Laplacian eigenvectors.

Graph convolutional network

Kipf and Welling [77] simplified this construction by further assuming $r \approx 2$ and $\lambda_n \approx 2$, resulting in filters of the form

$$\begin{aligned} g_\alpha(\mathbf{f}) &= \alpha_0 \mathbf{f} + \alpha_1(\Delta - \mathbf{I})\mathbf{f} \\ &= \alpha_0 \mathbf{f} - \alpha_1 \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}\mathbf{f}. \end{aligned} \tag{35}$$

Further constraining $\alpha = \alpha_0 = -\alpha_1$, one obtains filters represented by a single parameter,

$$g_\alpha(\mathbf{f}) = \alpha(\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})\mathbf{f}. \tag{36}$$

Because the eigenvalues of $\mathbf{I} + \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ are now in the range $[0, 2]$, repeated application of such a filter can result in numerical instability. This can be remedied by a renormalization

$$g_\alpha(\mathbf{f}) = \alpha\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}^{-1/2}\mathbf{f}, \tag{37}$$

where $\tilde{\mathbf{W}} = \mathbf{W} + \mathbf{I}$ and $\tilde{\mathbf{D}} = \mathrm{diag}\left(\sum_{j \neq i} \tilde{w}_{ij}\right)$.

Note that though we arrived at the constructions of ChebNet and graph convolutional network (GCN) starting in the spectral domain, they boil down to applying simple filters acting on the r- or one-hop neighborhood of the graph in the spatial domain. We consider these constructions to be examples of the more general graph neural network (GNN) framework.

GNN

GNNs [78] generalize the notion of applying the filtering operations directly on the graph via the graph weights. Similarly as Euclidean CNNs learn generic filters as linear combinations of localized, oriented bandpass and low-pass filters, a GNN learns at each layer a generic linear combination of graph low-pass and high-pass operators. These are given, respectively, by $f \mapsto \mathbf{W}f$ and $f \mapsto \Delta f$ and are thus generated by the degree matrix $\mathbf{D}$ and the diffusion matrix $\mathbf{W}$. Given a p-dimensional input signal on the vertices of the graph, represented by the $n \times p$ matrix $\mathbf{F}$, the GNN considers a generic nonlinear function $\eta_{\boldsymbol{\theta}}: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^q$, parameterized by trainable parameters $\boldsymbol{\theta}$ that is applied to all nodes of the graph,

$$\mathbf{g}_i = \eta_{\boldsymbol{\theta}}((\mathbf{W}\mathbf{f})_i, (\mathbf{D}\mathbf{f})_i). \tag{38}$$

In particular, choosing $\eta(\mathbf{a}, \mathbf{b}) = \mathbf{b} - \mathbf{a}$, one recovers the Laplacian operator $\Delta\mathbf{f}$, but more general, nonlinear choices for η yield trainable, task-specific diffusion operators. Similarly as with a CNN architecture, one can stack the resulting GNN layers $\mathbf{g} = C_{\boldsymbol{\theta}}(\mathbf{f})$ and interleave them with graph pooling operators. Chebyshev polynomials $T_r(\Delta)$ can be obtained with r layers of (38), making it possible, in principle, to consider ChebNet and GCN as particular instances of the GNN framework.

Historically, a version of GNN was the first formulation of deep learning on graphs, proposed in [49] and [78]. These works optimized over the parameterized steady state of some diffusion process (or random walk) on the graph. This can be interpreted as in (38) but using a large number of layers where each $C_{\boldsymbol{\theta}}$ is identical, as the forward propagation through the $C_{\boldsymbol{\theta}}$ approximate the steady state. Recent works [50], [51], [55], [79], [80] relax the requirements of approaching the steady state or using repeated applications of the same $C_{\boldsymbol{\theta}}$.

Because the communication at each layer is local to a vertex neighborhood, one may worry that it would take many layers to get information from one part of the graph to another, requiring multiple hops (this was one of the reasons for the use of the steady state in [78]). However, for many applications, it is not necessary for information to completely traverse

Three-Dimensional Shape Correspondence Application

Finding intrinsic correspondence between deformable shapes is a classical tough problem that underlies a broad range of vision and graphics applications, including texture mapping, animation, editing, and scene understanding [107]. From the machine-learning standpoint, correspondence can be thought of as a classification problem, where each point on the query shape is assigned to one of the points on a reference shape (serving as a label space) [108]. It is possible to learn the correspondence with a deep intrinsic network applied to some input feature vector $\mathbf{f}(x)$ at each point x of the query shape $\mathcal{X}$, producing an output $U_\Theta(\mathbf{f}(x))(y)$, which is interpreted as the conditional probability $p(y \mid x)$ of x being mapped to y [Figure S7(a)]. Using a training set of points with their ground-truth correspondence $\{x_i, y_i\}_{i \in \mathcal{I}}$, supervised learning is performed minimizing the multinomial regression loss

$$\min_\Theta - \sum_{i \in \mathcal{I}} \log U_\Theta(\mathbf{f}(x_i))(y_i) \tag{S14}$$

with respect to the network parameters Θ. The loss penalizes for the deviation of the predicted correspondence from the ground truth. We note that, while producing impressive results [Figure S7(b)], such an approach essentially learns pointwise correspondence, which then has to be postprocessed to satisfy certain properties, such as smoothness or bijectivity. Correspondence is an example of structured output, where the output of the network at one point depends on the output in other points (in the simplest setting, correspondence should be smooth, i.e., the output at nearby points should be similar) Litany et al. [109] proposed intrinsic structured prediction of shape correspondence by integrating a layer computing functional correspondence [106] into the deep neural network.

FIGURE S7. (a) The learning shape correspondence: an intrinsic deep network U_Θ is applied pointwise to some input features defined at each point. The output of the network at each point x of the query shape $\mathcal{X}$ is a probability distribution of the reference shape $\mathcal{Y}$ that can be thought of as a soft correspondence. (b) The intrinsic correspondence established between human shapes using intrinsic deep architecture (MoNet [54] with three convolutional layers). Signature of histogram orientations (SHOT) descriptors capturing the local normal vector orientations [110] were used in this example as input features. The correspondence is visualized by transferring texture from the leftmost reference shape. For additional examples, see [54].

the graph. Furthermore, note that the graphs at each layer of the network need not be the same. Thus, we can replace the original neighborhood structure with one's favorite multiscale coarsening of the input graph and operate on that to obtain the same flow of information as with the convolutional nets above (or rather more like a locally connected network [81]). This also allows producing a single output for the whole graph (for translation-invariant tasks), rather than a per-vertex output, by connecting each vertex to a special output node. Alternatively, one can allow η to use not only $\mathbf{Wf}$ and $\Delta\mathbf{f}$ at each node but also $\mathbf{W}^s\mathbf{f}$ for several diffusion scales $s > 1$ (as in [45]), giving the GNN the ability to learn algorithms like the power method and more directly accessing spectral properties of the graph. The GNN model can be further generalized to replicate other operators on graphs. For instance, the pointwise nonlinearity η can depend on the vertex type, allowing extremely rich architectures [50], [51], [55], [79], [80].

Charting-based methods

We now consider the second subclass of non-Euclidean learning problems, where we are given multiple domains. A prototypical application the reader should have in mind throughout this section is the problem of finding correspondence between shapes, modeled as manifolds (see "Three-Dimensional Shape Correspondence Application"). As we have seen, defining convolution in the spectral domain has an inherent drawback of the inability to adapt the model across different domains. We will therefore need to resort to an alternative generalization of the convolution in the spatial domain that does not suffer from this drawback.

Furthermore, note that in the setting of multiple domains, there is no immediate way to define a meaningful spatial pooling operation, as the number of points on different domains can vary, and their order can be arbitrary. It is, however, possible to pool pointwise features produced by a network by aggregating all the local information into a single vector. One possibility for such a pooling is computing the statistics of the pointwise features, e.g., the mean or covariance [47]. Note that after such a pooling, all of the spatial information is lost.

On a Euclidean domain, due to shift invariance the convolution can be thought of as passing a template at each point of the domain and recording the correlation of the template with the function at that point. Thinking of image filtering, this amounts to extracting a (typically square) patch of pixels, multiplying it elementwise with a template and summing up the results, then moving to the next position in a sliding window manner. Shift invariance implies that the very operation of extracting the patch at each position is always the same.

One of the major problems in applying the same paradigm to non-Euclidean domains is the lack of shift invariance, implying that the patch operator extracting a local patch would be position dependent. Furthermore, the typical lack of meaningful global parameterization for a graph or manifold forces to represent the patch in some local intrinsic system of coordinates. Such a mapping can be obtained by defining a set of weighting functions $v_1(x,\cdot),\ldots,v_J(x,\cdot)$ localized to positions near x (see examples in Figure 3). Extracting a patch amounts to averaging the function f at each point by these weights,

$$D_j(x)f = \int_X f(x')v_j(x,x')dx', j = 1,\ldots,J, \tag{39}$$

providing for a spatial definition of an intrinsic equivalent of convolution

$$(f \star g)(x) = \sum_j g_j D_j(x)f, \tag{40}$$

where g denotes the template coefficients applied on the patch extracted at each point. Overall, (39) and (40) act as a kind of nonlinear filtering of f, and the patch operator D is specified by defining the weighting functions $v_1,\ldots,v_J$. Such filters are localized by construction, and the number of parameters is equal to the number of weighting functions $J = O(1)$. Several frameworks for non-Euclidean CNNs essentially amount to different choices of these weights. The spectrum-free methods (ChebNet and GCN) described in the previous section can also be thought of in terms of local weighting functions, as it is easy to see the analogy between (40) and (34).

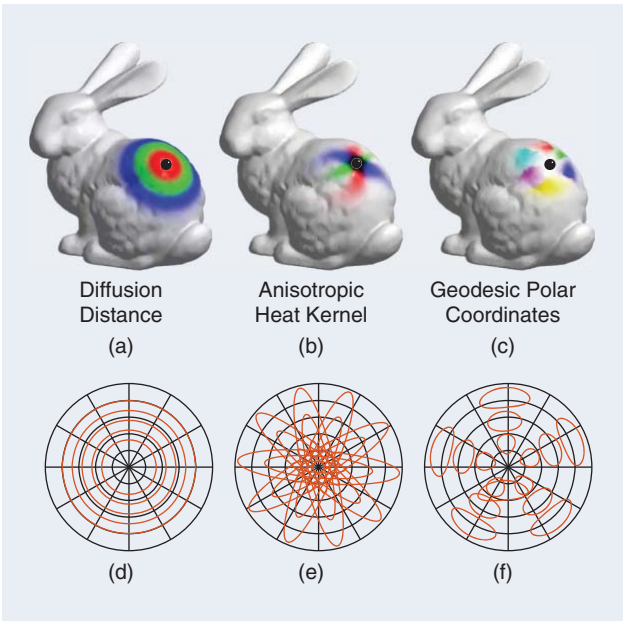


FIGURE 3. (a)–(c) The examples of intrinsic weighting functions used to construct a patch operator at the point marked in black (different colors represent different weighting functions). (a) Diffusion distance allows to map neighbor points according to their distance from the reference point, thus defining a 1-D system of local intrinsic coordinates. (b) Anisotropic heat kernels of different scale and orientations and (c) geodesic polar weights are 2-D systems of coordinates. (d)–(f) The representation of the weighting functions in the local polar (ρ,θ) system of coordinates (red curves represent the 0.5 level set).

Geodesic CNN

Because manifolds naturally come with a low-dimensional tangent space associated with each point, it is natural to work in a local system of coordinates in the tangent space [47]. In particular, on 2-D manifolds one can create a polar system of coordinates around x where the radial coordinate is given by some intrinsic distance $\rho(x') = d(x,x')$, and the angular coordinate $\theta(x)$ is obtained by ray shooting from a point at equispaced angles. The weighting functions in this case can be obtained as a product of Gaussians

$$v_{ij}(x,x') = e^{-(\rho(x')-\rho_i)^2/2\sigma_\rho^2} e^{-(\theta(x')-\theta_j)^2/2\sigma_\theta^2}, \tag{41}$$

where $i = 1,\ldots,J$ and $j = 1,\ldots,J'$ denote the indices of the radial and angular bins, respectively. The resulting JJ' weights are bins of width $\sigma_\rho \times \sigma_\theta$ in the polar coordinates [Figure 3(c) and (f)].

Anisotropic CNN

We have already seen the non-Euclidean heat equation (S5), whose heat kernel $h_t(x,\cdot)$ produces localized blob-like weights around the point x [see Figure S3(a)]. Varying the diffusion time t controls the spread of the kernel. However, such kernels are isotropic, meaning that the heat flows equally fast in all the directions. A more general anisotropic diffusion [48] equation on a manifold

$$f_t(x,t) = -\operatorname{div}(\mathbf{A}(x)\nabla f(x,t)) \tag{42}$$

involves the thermal conductivity tensor $\mathbf{A}(x)$ (in the case of 2-D manifolds, a 2×2 matrix is applied to the intrinsic gradient

in the tangent plane at each point), allowing modeling heat flow that is position and direction dependent [82]. A particular choice of the heat conductivity tensor proposed in [53] is

$$\mathbf{A}_{\alpha\theta}(x) = \mathbf{R}_\theta(x)\begin{pmatrix}\alpha & \\ & 1\end{pmatrix}\mathbf{R}_\theta^\top(x), \tag{43}$$

where the 2×2 matrix $\mathbf{R}_\theta(x)$ performs rotation of θ with respect to some reference (e.g., the maximum curvature) direction and $\alpha > 0$ is a parameter controlling the degree of anisotropy ($\alpha = 1$ corresponds to the classical isotropic case). The heat kernel of such anisotropic diffusion equation is given by the spectral expansion

$$h_{\alpha\theta t}(x,x') = \sum_{i\geq 0} e^{-t\lambda_{\alpha\theta i}}\phi_{\alpha\theta i}(x)\phi_{\alpha\theta i}(x'), \tag{44}$$

where $\phi_{\alpha\theta 0}(x), \phi_{\alpha\theta 1}(x), \ldots$ are the eigenfunctions and $\lambda_{\alpha\theta 0}, \lambda_{\alpha\theta 1}, \ldots$ the corresponding eigenvalues of the anisotropic Laplacian

$$\Delta_{\alpha\theta} f(x) = -\operatorname{div}(\mathbf{A}_{\alpha\theta}(x)\nabla f(x)). \tag{45}$$

The discretization of the anisotropic Laplacian is a modification of the cotangent formula (S12) on meshes or graph Laplacian (S9) on point clouds [48]. The anisotropic heat kernels $h_{\alpha\theta t}(x,\cdot)$ look like elongated rotated blobs [see Figure 3(b) and (e)], where the parameters α, θ and t control the elongation, orientation, and scale, respectively. Using such kernels as weighting functions v in the construction of the patch operator (39), it is possible to obtain a charting similar to the geodesic patches (roughly, θ plays the role of the angular coordinate and t of the radial one).

Mixture model network

Finally, as the most general construction of patches, Monti et al. [54] proposed defining at each point a local system of d-dimensional pseudocoordinates $\mathbf{u}(x,x')$ around x. On these coordinates, a set of parametric kernels $v_1(\mathbf{u}), \ldots, v_J(\mathbf{u})$ is applied, producing the weighting functions in (39). Rather than using fixed kernels, as in the previous constructions, Monti et al. use Gaussian kernels

$$v_j(\mathbf{u}) = \exp\Big(-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{u}-\boldsymbol{\mu}_j)\Big),$$

whose parameters ($d \times d$ covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J$ and $d \times 1$ mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J$) are learned [this choice allows interpreting intrinsic convolution (40) as a mixture of Gaussians, hence the name of the approach]. Learning not only the filters but also the patch operators in (40) affords additional DoF to the mixture model network (MoNet) architecture, which makes it currently the state-of-the-art approach in several applications. It is also easy to see that this approach generalizes the previous models, and, e.g., classical Euclidean CNNs as well as geodesic and anisotropic CNNs can be obtained as particular instances thereof [54]. MoNet can also be applied on general graphs using as the pseudocoordinates $\mathbf{u}$ some local graph features, such as vertex degree, geodesic distance, and so forth.

Combined spatial/spectral methods

The third alternative for constructing convolutionlike operations of non-Euclidean domains is jointly in spatial-frequency domain.

Windowed Fourier transform

One of the notable drawbacks of classical Fourier analysis is its lack of spatial localization. By virtue of the uncertainty principle, one of the fundamental properties of Fourier transforms, spatial localization comes at the expense of frequency localization and vice versa. In classical signal processing, this problem is remedied by localizing frequency analysis in a window $g(x)$, leading to the definition of the windowed Fourier transform (WFT, also known as *short-time Fourier transform* or *spectrogram* in signal processing),

$$(Sf)(x,\omega) = \int_{-\infty}^{\infty} f(x')\underbrace{g(x'-x)e^{-i\omega x'}}_{\overline{g_{x,\omega}(x')}}dx' \tag{46}$$

$$= \langle f, g_{x,\omega}\rangle_{L^2(\mathbb{R})}. \tag{47}$$

The WFT is a function of two variables: spatial location of the window x and the modulation frequency ω. The choice of the window function g allows control of the tradeoff between spatial and frequency localization (wider windows result in better frequency resolution). Note that WFT can be interpreted as inner products (47) of the function f with translated and modulated windows $g_{x,\omega}$, referred to as the WFT *atoms*.

The generalization of such a construction to non-Euclidean domains requires the definition of translation and modulation operators [83]. While modulation simply amounts to multiplication by a Laplacian eigenfunction, translation is not well defined due to the lack of shift invariance. It is possible to resort again to the spectral definition of a convolution-like operation (26), defining translation as convolution with a delta function,

$$(g \star \delta_{x'})(x) = \sum_{i\geq 0}\langle g,\phi_i\rangle_{L^2(\mathcal{X})}\langle \delta_{x'},\phi_i\rangle_{L^2(\mathcal{X})}\phi_i(x)$$

$$= \sum_{i\geq 0}\hat{g}_i\phi_i(x')\phi_i(x). \tag{48}$$

The translated and modulated atoms can be expressed as

$$g_{x',j}(x) = \phi_j(x')\sum_{i\geq 0}\hat{g}_i\phi_i(x)\phi_i(x'), \tag{49}$$

where the window is specified in the spectral domain by its Fourier coefficients $\hat{g}$. The WFT on non-Euclidean domains thus takes the form

$$(Sf)(x',j) = \langle f, g_{x',j}\rangle_{L^2(\mathcal{X})} = \sum_{i\geq 0}\hat{g}_i\phi_i(x')\langle f,\phi_i\phi_j\rangle_{L^2(\mathcal{X})}. \tag{50}$$

Due to the intrinsic nature of all the quantities involved in its definition, the WFT is also intrinsic.

Wavelets

Replacing the notion of frequency in time–frequency representations by that of scale leads to wavelet decompositions. Wavelets have been extensively studied in general graph domains [84]. Their objective is to define stable linear decompositions with atoms well localized both in space and frequency that can efficiently approximate signals with isolated singularities. Similarly to the Euclidean setting, wavelet families can be constructed either from spectral constraints or from spatial constraints.

The simplest of such families are Haar wavelets. Several bottom-up wavelet constructions on graphs were studied in [85] and [86]. In [87], the authors developed an unsupervised method that learns wavelet decompositions on graphs by optimizing a sparse reconstruction objective. In [88], ensembles of Haar wavelet decompositions were used to define deep wavelet scattering transforms on general domains, obtaining excellent numerical performance. Learning amounts to finding optimal pairings of nodes at each scale, which can be efficiently solved in polynomial time.

Localized SCNN

Boscaini et al. used the WFT as a way of constructing patch operators (39) on manifolds and point clouds and used in an intrinsic convolution-like construction (40). The WFT allows expressing a function around a point in the spectral domain in the form $D_j(x)f = (Sf)(x,j)$ [89]. Applying learnable filters to such patches (which in this case can be interpreted as spectral multipliers), it is possible to extract meaningful features that also appear to generalize across different domains. An additional DoF is the definition of the window, which can also be learned [89].

Applications

Network analysis

One of the classical examples used in many works on network analysis is citation networks. A citation network is a graph where vertices represent articles and there is a directed edge (i, j) if article i cites article j. Typically, vertex-wise features representing the content of the article (e.g., histogram of frequent terms in the article) are available. A prototypical classification application is to attribute each article to a field. Traditional approaches work vertex-wise, performing classification of each vertex's feature vector individually. More recently, it was shown that classification can be considerably improved using information from neighbor vertices, e.g., with a CNN on graphs [45], [77]. An example of the application of spectral and spatial graph CNN models on a citation network is shown in "Citation Network Analysis Application."

Another fundamental problem in network analysis is ranking and community detection. These can be estimated by solving an eigenvalue problem on an appropriately defined operator on the graph. For instance, the Fiedler vector (the eigenvector associated with the smallest nontrivial eigenvalue of the Laplacian) carries information on the graph partition with minimal cut [73], and the popular PageRank algorithm approximates page ranks with the principal eigenvector of a modified Laplacian operator. In some contexts, one may want develop data-driven versions of such algorithms that can adapt to model mismatch and perhaps provide a faster alternative to diagonalization methods. By unrolling power iterations, one obtains a GNN architecture whose parameters can be learned with backpropagation from labeled examples, similarly to the learned sparse coding paradigm [91]. We are currently exploring this connection by constructing multiscale versions of GNNs.

Recommender systems

Recommending movies on Netflix, friends on Facebook, or products on Amazon are a few examples of recommender systems that have recently become ubiquitous in a broad range of applications. Mathematically, a recommendation method can be posed as a matrix completion problem [92], where columns and rows represent users and items, respectively, and matrix values represent a score determining whether a user would like an item or not. Given a small subset of known elements of the matrix, the goal is to fill in the rest. A famous example is the Netflix challenge [93] offered in 2009 and carrying a US$1 million prize for the algorithm that can best predict user ratings for movies based on previous ratings. The size of the Netflix matrix is 480,000 movies × 18,000 users (8.5 billion elements), with only 0.011% known entries.

Recommending movies on Netflix, friends on Facebook, or products on Amazon are a few examples of recommender systems that have recently become ubiquitous in a broad range of applications. Such applications may benefit from geometric deep learning methods.

Several recent works proposed to incorporate geometric structure into matrix completion problems [94]–[97] in the form of column and row graphs representing similarity of users and items, respectively (see Figure 4). Such a geometric matrix completion setting makes meaningful, e.g., the notion of smoothness of the matrix values and was shown beneficial for the performance of recommender systems.

In a recent work, Monti et al. [56] proposed addressing the geometric matrix completion problem by means of a learnable model combining a multigraph CNN (MGCNN) and a recurrent neural network (RNN). Multigraph convolution can be thought of as a generalization of the standard bidimensional image convolution, where the domains of the rows and the columns are now different (in our case, user and item graphs). The features extracted from the score matrix by means of the MGCNN are then passed to an RNN, which produces a sequence of incremental updates of the score values. Overall, the model can be considered as a learnable diffusion of the scores, with the main advantage compared to traditional approach being a fixed number of variables independent of the matrix size. The MGCNN achieved

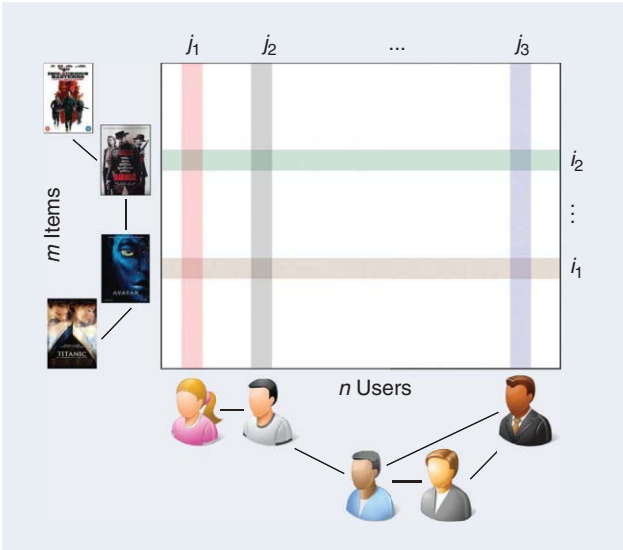


FIGURE 4. The geometric matrix completion exemplified on the famous Netflix movie recommendation problem. The column and row graphs represent the relationships between users and items, respectively.

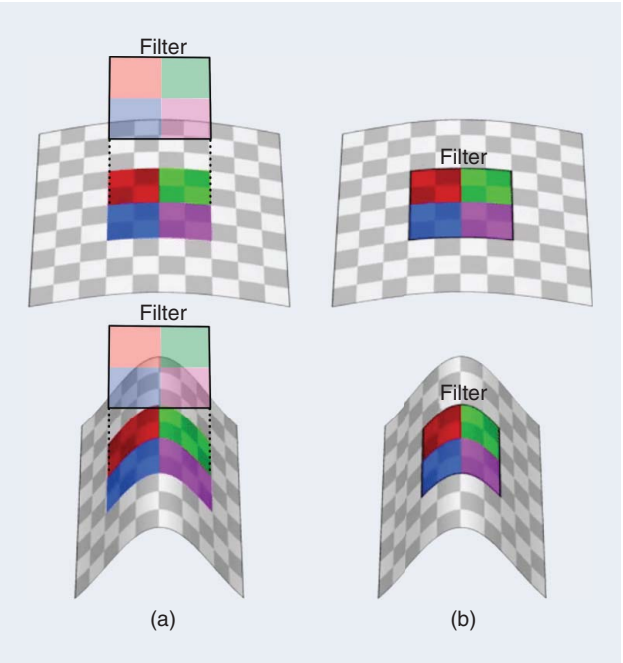


FIGURE 5. An illustration of the difference between (a) classical CNN applied to a 3-D shape (checkered surface) considered as a Euclidean object and (b) a geometric CNN applied intrinsically on the surface. In the latter case, the convolutional filters (visualized as a colored window) are deformation invariant by construction.

state-of-the-art results on several classical matrix completion challenges and, on a more conceptual level, could be a very interesting practical application of geometric deep learning to a classical signal processing problem of matrix completion.

Computer vision and graphics

The computer-vision community has recently shown an increasing interest in working with 3-D geometric data, mainly due to the emergence of affordable range-sensing technology, such as Microsoft Kinect or Intel RealSense. Many machine-learning techniques successfully working on images were tried as is on 3-D geometric data, represented for this purpose in some way digestible by standard frameworks, e.g., as range images [98], [99] or rasterized volumes [100], [101]. The main drawback of such approaches is their treatment of geometric data as Euclidean structures. First, for complex 3-D objects, Euclidean representations, such as depth images or voxels, may lose significant parts of the object or its fine details or even break its topological structure. Second, Euclidean representations are not intrinsic and vary when changing pose or deforming the object. Achieving invariance to shape deformations, a common requirement in many vision applications, demands very complex models and huge training sets due to the large number of DoF involved in describing nonrigid deformations [see Figure 5(a)].

The computer-vision community has recently shown an increasing interest in working with 3-D geometric data.

In the domain of computer graphics, on the other hand, working intrinsically with geometric shapes is a standard practice. In this field, 3-D shapes are typically modeled as Riemannian manifolds and are discretized as meshes. Numerous studies (see, e.g., [102]–[106]) have been devoted to designing local and global features, e.g., for establishing similarity or correspondence between deformable shapes with guaranteed invariance to isometries.

However, different applications in computer vision and graphics may require completely different features. For instance, to establish feature-based correspondence between a collection of human shapes, one would desire the descriptors of corresponding anatomical parts (e.g., noses, mouths) to be as similar as possible across the collection (see Figure 6(a)). In other words, such descriptors should be invariant to the collection variability. Conversely, for shape classification, one would like descriptors that emphasize the subject-specific characteristics and, e.g., distinguish between two different nose shapes (see Figure 6b). Deciding a priori which structures should be used and which should be ignored is often hard or sometimes even impossible. Moreover, axiomatic modeling of geometric noise, such as 3-D scanning artifacts, turns out to be extremely hard.

By resorting to intrinsic deep neural networks on manifolds, the invariance to isometric deformations is automatically built into the model, thus vastly reducing the number of DoF required to describe the invariance class. Roughly speaking, the intrinsic deep model will try to learn residual deformations that deviate from the isometric model. Geometric deep learning can be applied to several problems in 3-D shape analysis, which can be divided into two classes. First are problems like local descriptor learning [47], [53] or correspondence learning [48] (see the example in "Three-Dimensional

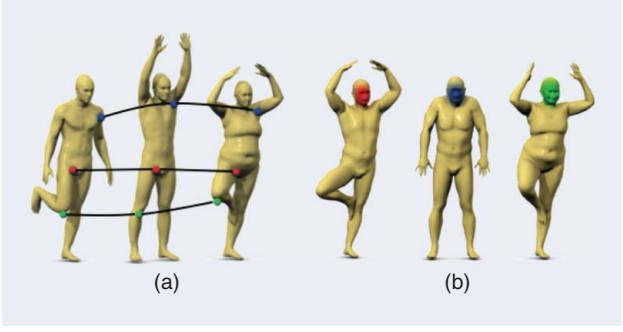


FIGURE 6. (a) The features used for shape correspondence should ideally manifest invariance across the shape class (e.g., the knee feature shown here should not depend on the specific person). (b) The features used for shape retrieval, on the contrary, should be specific to a shape within the class to allow distinguishing between different people. Similar features are marked with the same color. Handcrafting the right feature for each application is a very challenging task.

Shape Correspondence Application"), in which the output of the network is pointwise. The inputs to the network are some pointwise features, e.g., color texture or simple geometric features, such as normals. Using a CNN architecture with multiple intrinsic convolutional layers, it is possible to produce nonlocal features that capture the context around each point. The second type of problems, such as shape recognition, require the network to produce a global shape descriptor, aggregating all the local information into a single vector using, e.g., the covariance pooling [47].

The recent emergence of geometric deep-learning methods in various communities and application domains allows us to proclaim that we might be witnessing a new field being born.

Particle physics and chemistry

Many areas of experimental science are interested in studying systems of discrete particles defined over a low-dimensional phase space. For instance, the chemical properties of a molecule are determined by the relative positions of its atoms, and the classification of events in particle accelerators depends upon position, momentum, and spin of all the particles involved in the collision.

The behavior of an N-particle system is ultimately derived from solutions of the Schrödinger equation, but its exact solution involves diagonalizing a linear system of exponential size. In this context, an important question is whether one can approximate the dynamics with a tractable model that incorporates by construction the geometric stability postulated by the Schrödinger equation and at the same time has enough flexibility to adapt to data-driven scenarios and capture complex interactions.

An instance l of an N_l-particle system can be expressed as

$$f_l(t) = \sum_{j=1}^{N_l} \alpha_{j,l}\delta(t - x_{j,l}),$$

where $(\alpha_{j,l})$ model particle-specific information, such as the spin, and $(x_{j,l})$ are the locations of the particles in a given phase space. Such a system can be recast as a signal defined over a graph with $|\mathcal{V}_l| = N_l$ vertices and edge weights $\mathbf{W}_l = (\phi(\alpha_{i,l}, \alpha_{j,l}, x_{i,l}, x_{j,l}))$ expressed through a similarity kernel capturing the appropriate priors. GNNs are currently being applied to perform event classification, energy regression, and anomaly detection in high-energy physics experiments, such as the Large Hadron Collider, and neutrino detection in the IceCube Observatory. Recently, models based on GNNs have been applied to predict the dynamics of N-body systems [111], [112], showing excellent prediction performance.

Molecule design

A key problem in material and drug design is predicting the physical, chemical, or biological properties (such as solubility of toxicity) of a novel molecule from its structure. State-of-the-art methods rely on hand-crafted molecule descriptors, such as circular fingerprints [113]–[115]. A recent work from Harvard University in Cambridge, Massachusetts [55] proposed modeling molecules as graphs (where vertices represent atoms and edges represent chemical bonds) and employing GCNNs to learn the desired molecule properties. The authors' approach has significantly outperformed handcrafted features. This work opens a new avenue in molecule design that might revolutionize the field.

Medical imaging

An application area where signals are naturally collected on non-Euclidean domains and where the methodologies we reviewed could be very useful is brain imaging. A recent trend in neuroscience is to associate functional magnetic resonance imaging traces with a precomputed connectivity rather than inferring it from the traces themselves [116]. In this case, the challenge consists in processing and analyzing an array of signals collected over a complex topology, which results in subtle dependencies. For example, in a recent work from Imperial College London [117], GCNNs were used to detect disruptions of the brain functional networks associated with autism.

Open problems and future directions

The recent emergence of geometric deep-learning methods in various communities and application domains, which we tried to overview in this article, allows us to proclaim, perhaps with some caution, that we might be witnessing a new field being born. We expect the following years to bring exciting new methods and applications, and conclude our review with a few observations of current key difficulties and potential directions of future research.

Many disciplines dealing with geometric data employ some empirical models or handcrafted features. This is a typical situation in geometry processing and computer graphics, where axiomatically constructed features are used to analyze 3-D shapes, or computational sociology, where it is common

to first come up with a hypothesis and then test it on the data [22]. Yet, such models assume some prior knowledge (e.g., isometric shape deformation model) and often fail to correctly capture the full complexity and richness of the data. In computer vision, departing from handcrafted features toward generic models learnable from the data in a task-specific manner has brought a breakthrough in performance and led to an overwhelming trend in the community to favor deep-learning methods. Such a shift has not occurred yet in the fields dealing with geometric data due to the lack of adequate methods, but there are the first indications of a coming paradigm shift.

Generalization

Generalizing deep-learning models to geometric data requires not only finding non-Euclidean counterparts of basic building blocks (such as convolutional and pooling layers) but also generalization across different domains. Generalization capability is a key requirement in many applications, including computer graphics, where a model is learned on a training set of non-Euclidean domains (3-D shapes) and then applied to previously unseen ones. Spectral formulation of convolution allows designing CNNs on a graph, but the model learned this way on one graph cannot be straightforwardly applied to another one, because the spectral representation of convolution is domain dependent. A possible remedy to the generalization problem of spectral methods is the recent architecture proposed in [118], applying the idea of spatial transformer networks [119] in the spectral domain. This approach is reminiscent of the construction of compatible orthogonal bases by means of joint Laplacian diagonalization [75], which can be interpreted as an alignment of two Laplacian eigenbases in a k-dimensional space.

The spatial methods, on the other hand, allow generalization across different domains, but the construction of low-dimensional local spatial coordinates on graphs turns out to be rather challenging. In particular, the construction of anisotropic diffusion on general graphs is an interesting research direction. The spectrum-free approaches also allow generalization across graphs, at least in terms of their functional form. However, if multiple layers of (38) are used with no nonlinearity or learned parameters θ, simulating a high power of the diffusion, the model may behave differently on different kinds of graphs. Understanding under what circumstances and to what extent these methods generalize across graphs is currently being studied.

Time-varying domains

An interesting extension of geometric deep-learning problems discussed in this review is coping with signals defined over a dynamically changing structure. In this case, we cannot assume a fixed domain and must track how these changes affect signals. This could prove useful to tackle applications like abnormal activity detection in social or financial networks. In the domain of computer graphics and vision, potential applications deal with dynamic shapes (e.g., 3-D video captured by a range sensor).

Directed graphs

Dealing with directed graphs is also a challenging topic, as such graphs typically have nonsymmetric Laplacian matrices that do not have orthogonal eigendecomposition allowing easily interpretable spectral-domain constructions. Citation networks, which are directed graphs, are often treated as undirected graphs (including in our example in "Three-Dimensional Shape Correspondence Application") considering citations between two articles without distinguishing which article cites which. This obviously may lose important information.

Synthesis problems

Our main focus in this review was primarily on analysis problems on non-Euclidean domains. Not less important is the question of data synthesis. There have been several recent attempts to try to learn a generative model allowing to synthesize new images [120] and speech waveforms [121]. Extending such methods to the geometric setting seems a promising direction, though the key difficulty is the need to reconstruct the geometric structure (e.g., an embedding of a 2-D manifold in the 3-D Euclidean space modeling a deformable shape) from some intrinsic representation [122].

One of the main reasons for the computational efficiency of deep-learning architectures is the assumption of regularly structured data on a 1-D or 2-D grid.

Computation

The final consideration is a computational one. All existing deep-learning software frameworks are primarily optimized for Euclidean data. One of the main reasons for the computational efficiency of deep-learning architectures (and one of the factors that contributed to their renaissance) is the assumption of regularly structured data on a 1-D or 2-D grid, allowing to take advantage of modern GPU hardware. Geometric data, on the other hand, in most cases do not have a grid structure, requiring different ways to achieve efficient computations. It seems that computational paradigms developed for large-scale graph processing are more adequate frameworks for such applications.

Acknowledgments

We are grateful to Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Xavier Bresson, Thomas Kipf, and Michaël Defferard for their comments and for providing some of the figures used in this article. This work was supported in part by the European Research Council grant numbers 307047 (COMET) and 724228 (LEMAN), a Google Faculty Research Award, a Radcliffe fellowship, a Rudolf Diesel fellowship, and Nvidia equipment grants.

Authors

Michael M. Bronstein (michael.bronstein@usi.ch) received his B.S. degree in electrical engineering and his Ph.D. degree in computer science from the Technion Israel Institute of Technology, Haifa, in 2002 and 2007, respectively. He is currently an associate professor at the University of Lugano, Switzerland, and Tel Aviv University, Israel. He also serves as a principal engineer at Intel Perceptual Computing, Israel. His main research interest is in theoretical and computational methods for geometric data analysis. He was selected as an ACM Distinguished Speaker, a World Economic Forum Young Scientist, and a member of the Young Academy of Europe. He received three ERC grants, a Radcliffe fellowship from Harvard University, a Rudolf Diesel fellowship from the Technical University of Munich, and a Google faculty award.

Joan Bruna (bruna@cims.nyu.edu) received his B.S. degree in mathematics and electrical engineering in 2002 from the Universitat Politècnica de Catalunya, Barcelona, Spain, his M.S. degree in applied mathematics in 2005 from the École normale supérieure Cachan, France, and his Ph.D. degree in applied mathematics in 2013 from the École Polytechnique, Palaiseau, France. He was a postdoctoral researcher at the Courant Institute, New York University (NYU), and a postdoctoral fellow at Facebook AI Research, Menlo Park, California. In 2015, he became an assistant professor at the University of California, Berkeley Statistics Department, and, starting in the fall of 2016, he joined the Courant Institute, NYU, as an assistant professor in computer science, data science, and mathematics (affiliated). His research interests include invariant signal representations, high-dimensional statistics and stochastic processes, and deep learning and its applications to signal processing.

Yann LeCun (yann@fb.com) received his electrical engineering diploma in 1983 from the Ecole Superieure d'Ingenieurs en Electrotechnique et Electronique, Paris, and his Ph.D. degree in computer science from the Universitè Pierre et Marie Curie, Paris, France, in 1987. After postdoctoral research at the University of Toronto, he joined AT&T Bell Labs in Holmdel, New Jersey, in 1988, where he became head of the Image Processing Research Department. He joined New York University (NYU) as a professor in 2003 after a brief period as a fellow of the NEC Research Institute in Princeton, New Jersey. He is currently the director of artificial intelligence research at Facebook and a professor at NYU. Since the late 1980s, he has been working on deep-learning methods, particularly the convolutional network model. He has been on the editorial board of *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Neural Networks*. He received the IEEE Neural Network Pioneer Award in 2014.

Arthur Szlam (aszlam@fb.com) received his Ph.D. degree in computer science from Yale University, New Haven, Connecticut. He is a research scientist at Facebook AI Research, Menlo Park, California. Prior to joining Facebook, he was a postdoctoral fellow at the Institute for Mathematics and Its Applications at the University of Minnesota, Minneapolis, and the Institute for Pure and Applied Mathematics, University of California, Los Angeles, and an assistant professor at City College of New York. He is a Sloan Foundation fellow. His research interests are in computational harmonic analysis, the relationships between smoothness, frequency, and scale on graphs and data clouds, and applications to signal processing and machine learning.

Pierre Vandergheynst (pierre.vandergheynst@epfl.ch) received his M.S. degree in theoretical physics in 1994 and his Ph.D. degree in mathematical physics from the Université catholique de Louvain, Belgium, in 1998. He is a full professor at the Ecole Polytechnique Federale de Lausanne, Switzerland. He has authored or coauthored more than 50 journal papers, one monograph, and several book chapters. He has been an associate editor of *IEEE Transactions on Signal Processing* since 2007, a member of technical committees for various conferences, and general cochair of the 2008 European Signal Processing Conference. He is a laureate of the Apple Research and Technology Support Award and of the 2010–2011 De Boelpaepe prize from the Royal Academy of Sciences of Belgium. He is strongly involved in technology transfer, having cofounded two startups, and holds numerous patents. His research focuses on harmonic analysis, sparse approximations, and mathematical image processing with applications to higher-dimensional, complex data processing.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černockỳ, "Strategies for training large scale neural network language models," in *Proc. 2011 IEEE Workshop Automatic Speech Recognition and Understanding*, pp. 196–201.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2014, pp. 3104–3112.

[5] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2010, pp. 253–256.

[6] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2011, pp. 253–256.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

[10] K. Simonyan and A. Zisserman. (2014). Very deep convolutional networks for large-scale image recognition. [Online]. Available: https://arXiv:1409.1556

[11] K. He, X. Zhang, S. Ren, and J. Sun. (2015). Deep residual learning for image recognition. [Online]. Available: https://arXiv:1512.03385

[12] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[14] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[15] D. J. Field, "What the statistics of natural images tell us about visual coding," in *Proc. SPIE-Int. Society for Optical Engineering*, vol. 1077, p. 269, 1989.

[16] P. Mehta and D. J. Schwab. (2014). An exact mapping between the variational renormalization group and deep learning. [Online]. Available: https://arXiv:1410.3831

[17] S. Mallat, "Group invariant scattering," *Commun. Pure and Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[18] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, 2013.

[19] M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, "A mathematical motivation for complex-valued convolutional networks," *Neural Computation*, vol. 28, no. 5, pp. 815–825, 2016.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten ZIP code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[21] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. (2013). Maxout networks. [Online]. Available: https://arXiv:1302.4389

[22] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Life in the network: The coming age of computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[23] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, M. J. Schilstra, P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri, "A genomic regulatory network for development," *Science*, vol. 295, no. 5560, pp. 1669–1678, 2002.

[24] M. B. Wakin, D. L. Donoho, H. Choi, and R. G. Baraniuk, "The multiscale structure of non-differentiable image manifolds," in *Proc. SPIE-Int. Society for Optical Engineering*, vol. 5914, p. 59141B, 2005.

[25] N. Verma, S. Kpotufe, and S. Dasgupta, "Which spatial partition trees are adaptive to intrinsic dimension?" in *Proc. Uncertainty in Artificial Intelligence*, 2009, pp. 565–574.

[26] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[27] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[28] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, pp. 2579–2605, 2008.

[29] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[30] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[31] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1735–1742.

[32] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2014, pp. 701–710.

[33] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *Proc. Int. World Wide Web Conf. (WWW)*, 2015, pp. 1067–1077.

[34] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *Proc. Int. Conf. Information and Knowledge Management (CIKM)*, 2015, pp. 891–900.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). Efficient estimation of word representations in vector space. [Online]. Available: https://arXiv:1301.3781

[36] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[37] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, pp. 177–183, 2007.

[38] J. Sun, M. Ovsjanikov, and L. J. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.

[39] R. Litman and A. M. Bronstein, "Learning spectral descriptors for deformable shape correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 171–180, 2014.

[40] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3, pp. 75–174, 2010.

[41] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2013, pp. 3111–3119.

[42] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2011, pp.1082–1090.

[43] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[44] M. Henaff, J. Bruna, and Y. LeCun. (2015). Deep convolutional networks on graph-structured data. [Online]. Available: https://arXiv:1506.05163

[45] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2016, pp. 3844–3852.

[46] J. Atwood and D. Towsley. (2016). Diffusion-convolutional neural networks. [Online]. Available: https://arXiv:1511.02136v2

[47] J. Masci, D. Boscaini, M. M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on Riemannian manifolds," in *Proc. Int. IEEE Workshop 3-D Representation and Recognition (3DRR)*, 2015, pp. 832–840.

[48] D. Boscaini, J. Masci, E. Rodolà, and M. M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2016, pp. 3189–3197.

[49] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2005, pp. 729–734.

[50] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. (2015). Gated graph sequence neural networks. [Online]. Available: https://arXiv:1511.05493

[51] S. Sukhbaatar, A. Szlam, and R. Fergus. (2016). Learning multiagent communication with backpropagation. [Online]. Available: https://arXiv:1605.07736

[52] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. (2013). Spectral networks and locally connected networks on graphs. [Online.] Available: https://arXiv.1312.6203

[53] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, and D. Cremers, "Anisotropic diffusion descriptors," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 431–441, 2016.

[54] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, (2017). Geometric deep learning on graphs and manifolds using mixture model CNNs. [Online.] Available: https://arXiv.1611.08402

[55] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2015, pp. 2224–2232.

[56] F. Monti, X. Bresson, and M. M. Bronstein. (2017). Geometric matrix completion with recurrent multi-graph neural networks. [Online]. Available: https://arXiv.1704.06803

[57] S. Mallat, "Understanding deep convolutional networks," *Philos. Trans. R. Soc. London A, Math. Phys. Sci.*, vol. 374, no. 2065, 2016. doi: 10.1098/rsta.2015.0207.

[58] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[59] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox "Flownet: Learning optical flow with convolutional networks," in *Proc. 2015 Int. Conf. Computer Vision*, 2015, pp. 2758–2766.

[60] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. (2014). Striving for simplicity: The all convolutional net. [Online]. Available: https://arXiv:1412.6806

[61] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1999.

[62] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proc. 18th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 192–204.

[63] I. Safran and O. Shamir. (2015). On the quality of the initial basin in overspecified neural networks. [Online]. Available: https://arXiv:1511.04210

[64] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2016, pp. 586–594.

[65] T. Chen, I. Goodfellow, and J. Shlens. (2015). Net2net: Accelerating learning via knowledge transfer. [Online]. Available: https://arXiv:1511.05641

[66] C. D. Freeman and J. Bruna. (2017). Topology and geometry of half-rectified network optimization. [Online]. Available: https://arXiv.1611.01540

[67] J. Nash, "The imbedding problem for Riemannian manifolds," *Ann. Math.*, vol. 63, no. 1, pp. 20–63, 1956.

[68] M. Wardetzky, S. Mathur, F. Kälberer, and E. Grinspun, "Discrete Laplace operators: No free lunch," in *Proc. 5th Eurographics Symp. Geometry Processing (SGP)*, 2007, pp. 33–37.

[69] M. Wardetzky, "Convergence of the cotangent formula: An overview," in *Discrete Differential Geometry*. Cambridge, MA: Birkhäuser, 2008, pp. 275–286.

[70] U. Pinkall and K. Polthier, "Computing discrete minimal surfaces and their conjugates," *Exp. Mathematics*, vol. 2, no. 1, pp. 15–36, 1993.

[71] S. Rosenberg, *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[72] L.-H. Lim. (2015). Hodge Laplacians on graphs. [Online]. Available: https://arXiv:1507.05379

[73] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[74] A. Kovnatsky, M. M. Bronstein, A. M. Bronstein, K. Glashoff, and R. Kimmel, "Coupled quasi-harmonic bases," *Comput. Graph. Forum*, vol. 32, no. 2, pp. 439–448, 2013.

[75] D. Eynard, A. Kovnatsky, M. M. Bronstein, K. Glashoff, and A. M. Bronstein, "Multimodal manifold analysis by simultaneous diagonalization of Laplacians," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2505–2517, 2015.

[76] N. L. Roux, Y. Bengio, P. Lamblin, M. Joliveau, and B. Kégl, "Learning the 2-d topology of images," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2008, pp. 841–848.

[77] T. N. Kipf and M. Welling. (2016). Semi-supervised classification with graph convolutional networks. [Online]. Available: https://arXiv:1609.02907

[78] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, 2009.

[79] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. (2016). A compositional object-based approach to learning physical dynamics. [Online]. Available: https://arXiv:1612.00341

[80] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, and K. Kavukcuoglu, "Interaction networks for learning about objects, relations and physics," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2016, pp. 4502–4510.

[81] A. Coates and A. Y. Ng, "Selecting receptive fields in deep networks," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2011, pp. 2528–2536.

[82] M. Andreux, E. Rodolà, M. Aubry, and D. Cremers, "Anisotropic Laplace-Beltrami operators for shape analysis," in *Proc. 6th Workshop Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA)*, 2014, pp. 299–312.

[83] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 260–291, 2016.

[84] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Appl. Comput. Harmonic Anal.*, vol. 21, no. 1, pp. 53–94, 2006.

[85] A. D. Szlam, M. Maggioni, R. R. Coifman, and J. C. Bremer, Jr., "Diffusion-driven multiscale analysis on manifolds and graphs: Top-down and bottom-up constructions," *Proc. SPIE-Int. Society for Optical Engineering*, vol. 5914, p. 59141D, 2005.

[86] M. Gavish, B. Nadler, and R. R. Coifman, "Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning," in *Proc. 27th Int. Conf. Machine Learning (ICML-10)*, 2010, pp. 367–374.

[87] R. Rustamov and L. J. Guibas, "Wavelets on graphs via deep learning," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2013, pp. 998–1006.

[88] X. Cheng, X. Chen, and S. Mallat, "Deep Haar scattering networks," *Information and Inference*, vol. 5, pp. 105–133, 2016.

[89] D. Boscaini, J. Masci, S. Melzi, M. M. Bronstein, U. Castellani, and P. Vandergheynst, "Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks," *Comput. Graph. Forum*, vol. 34, no. 5, pp. 13–23, 2015.

[90] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, p. 93, 2008.

[91] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Machine Learning (ICML-10)*, 2010, pp. 399–406.

[92] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[93] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[94] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King, "Recommender systems with social regularization," in *Proc. Web Search and Data Mining*, 2011, pp. 287–296.

[95] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst. (2014). Matrix completion on graphs. [Online]. Available: https://arXiv:1408.1717

[96] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2015, pp. 2107–2115.

[97] D. Kuang, Z. Shi, S. Osher, and A. Bertozzi. (2016). A harmonic extension approach for collaborative ranking. [Online]. Available: https://arXiv:1602.05127

[98] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2015, pp. 945–953.

[99] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li, "Dense human body correspondences using convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1544–1553.

[100] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D shapenets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[101] C. R. Qi, H. Su, M. Niessner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5648–5656.

[102] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 5, pp. 1168–1172, 2006.

[103] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1704–1711.

[104] V. Kim, Y. Lipman, and T. Funkhouser, "Blended intrinsic maps," *ACM Trans. Graph.*, vol. 30, no. 4, p. 79, 2011.

[105] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov, "ShapeGoogle: Geometric words and expressions for invariant shape retrieval," *ACM Trans. Graph.*, vol. 30, no. 1, p. 1, 2011.

[106] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. J. Guibas, "Functional maps: A flexible representation of maps between shapes," *ACM Trans. Graph.*, vol. 31, no. 4, p. 30, 2012.

[107] S. Biasotti, A. Cerri, A. M. Bronstein, and M. M. Bronstein, "Recent trends, applications, and perspectives in 3D shape similarity assessment," *Comput. Graph. Forum*, vol. 35, no. 6, pp. 87–119, 2016.

[108] E. Rodolà, S. Rota Bulo, T. Windheuser, M. Vestner, and D. Cremers, "Dense non-rigid shape correspondence using random forests," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 4177–4184.

[109] O. Litany, E. Rodolà, A. M. Bronstein, and M. M. Bronstein. (2017). Deep functional maps: Structured prediction for dense shape correspondence. [Online]. Available: https://arXiv.1704.08686

[110] F. Tombari, S. Salti, and L. D Stefano, "Unique signatures of histograms for local surface description," in *Proc. European Conf. Computer Vision (ECCV)*, 2010, pp. 356–369.

[111] P. W. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu, "Interaction networks for learning about objects, relations and physics," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 4502–4510.

[112] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. (2016). A compositional object-based approach to learning physical dynamics. [Online]. Available: https://arXiv.1612.00341

[113] H. L. Morgan, "The generation of a unique machine description for chemical structure," *J. Chem. Documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[114] R. C. Glem, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, and J. Smith, "The generation of a unique machine description for chemical structure," *Investigational Drugs*, vol. 9, no. 3, pp. 199–204, 2006.

[115] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inform. and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[116] M. G. Preti, T. A. Bolton, and D. Van De Ville. (2016). The dynamic functional connectome: State-of-the-art and perspectives,. *Science*. [Online] http://www.sciencedirect.com/science/article/pii/S1053811916307881

[117] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert. (2017). Distance metric learning using graph convolutional networks: Application to functional brain networks. [Online]. Available: https://arXiv:1703.02161

[118] L. Yi, H. Su, X. Guo, and L. J. Guibas. (2017). SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation. [Online]. Available: https://arXiv.1612.00606

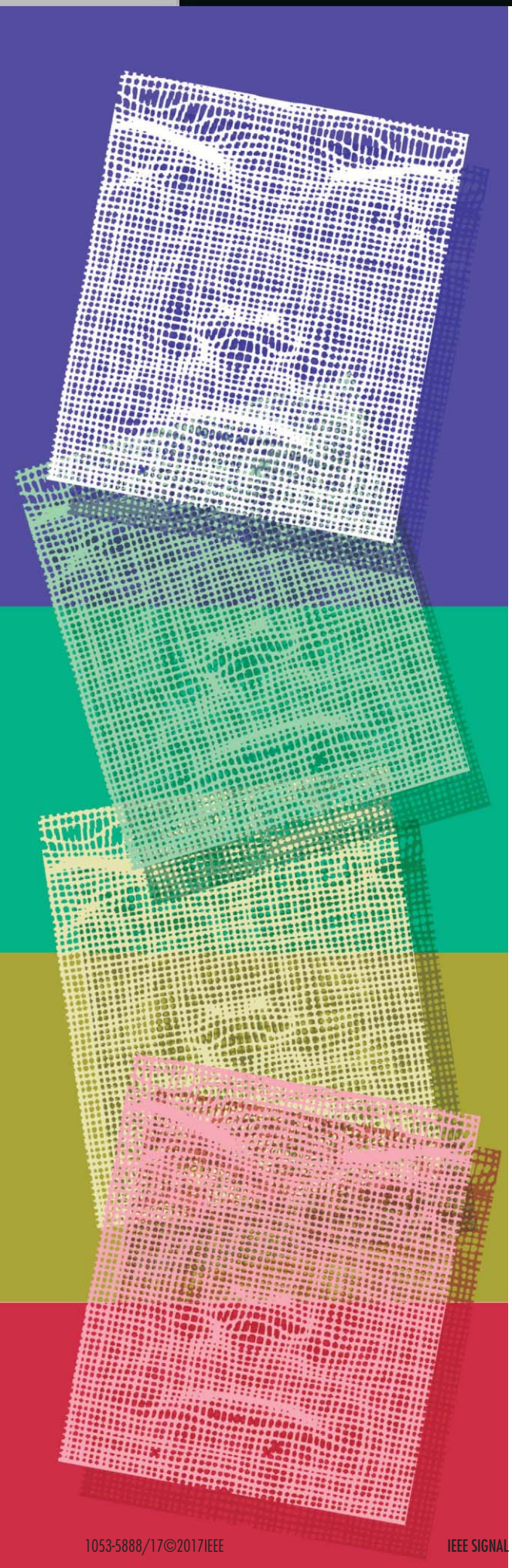
[119] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Int. Conf. Neural Information Processing Systems (NIPS)*, 2015, pp. 2017–2025.

[120] A. Dosovitskiy, J. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1538–1546.

[121] S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. (2016). Wavenet: A generative model for raw audio. [Online]. Available: https://arXiv:1609.03499

[122] D. Boscaini, D. Eynard, D. Kourounis, and M. M. Bronstein, "Shape-from-operator: Recovering shapes from intrinsic operators," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 265–274, 2015.

SP



Soheil Kolouri, Se Rim Park, Matthew Thorpe,
Dejan Slepčev, and Gustavo K. Rohde

Optimal Mass Transport

Signal processing and machine-learning applications

Transport-based techniques for signal and data analysis have recently received increased interest. Given their ability to provide accurate generative models for signal intensities and other data distributions, they have been used in a variety of applications, including content-based retrieval, cancer detection, image superresolution, and statistical machine learning, to name a few, and they have been shown to produce state-of-the-art results. Moreover, the geometric characteristics of transport-related metrics have inspired new kinds of algorithms for interpreting the meaning of data distributions. Here, we provide a practical overview of the mathematical underpinnings of mass transport-related methods, including numerical implementation, as well as a review, with demonstrations, of several applications. Software accompanying this article is available from [43].

Purposes for optimal mass transport

Motivation and goals

Numerous applications in science and technology depend on effective modeling and information extraction from signal and image data. Examples include being able to distinguish between benign and malignant tumors in medical images; learning models (e.g., dictionaries) for solving inverse problems; identifying people from images of faces, voice profiles, or fingerprints; and many others. Techniques based on the mathematics of optimal mass transport, also known as *Earth Mover's Distance* in engineering-related fields, have received significant attention recently given their ability to incorporate spatial (in addition to intensity) information when comparing signals, images, and other data sources, thus giving rise to different geometric interpretations of data distributions. These techniques have been used to simplify and augment the accuracy of numerous pattern

Digital Object Identifier 10.1109/MSP.2017.2695801
Date of publication: 11 July 2017

recognition-related problems. Some examples covered in this article include image retrieval [32], [44], signal and image representation [25], [27], [40], [50], inverse problems [30], cancer detection [4], [39], texture and color modeling [18], [41], shape and image registration [22], [29], and machine learning [12], [17], [19], [28], [36], [42], to name a few. This article is meant to serve as an introductory guide to those wishing to familiarize themselves with these emerging techniques. Specifically, we

- provide a brief overview of key mathematical concepts related to optimal mass transport
- describe recent advances in transport-related methodology and theory
- provide a practical overview of their applications in modern signal analysis, modeling, and learning problems.

Why transport?

In recent years, numerous techniques for signal and image analysis have been developed to address important learning and estimation problems. Researchers working to unveil solutions to these problems have found it necessary to develop techniques to compare signal intensities across different signal/image coordinates. A common problem in medical imaging, for example, is the analysis of magnetic resonance images with the goal of learning about brain morphology differences between healthy and diseased populations. Decades of research in this area have culminated with techniques such as voxel- and deformation-based morphology that make use of nonlinear registration methods to understand differences in tissue density and locations. Likewise, the development of dynamic time-warping techniques was necessary to enable the comparison of time series data more meaningfully without confounds from commonly encountered variations in time. Furthermore, researchers desiring to create realistic models of facial appearance have long understood that appearance models for the eyes, lips, nose, and other facial features are significantly different and thus must be dependent on a position relative to a fixed anatomy. The pervasive success of these as well as other techniques, such as optical flow, level-set methods, and deep neural networks, have shown that 1) nonlinearity and 2) modeling the location of pixel intensities are essential concepts to keep in mind when solving modern regression problems related to estimation and classification.

The previously mentioned methodology for modeling appearance and learning morphology, time series analysis and predictive modeling, deep neural networks for classification of sensor data, and the like is algorithmic in nature. The transport-related techniques reviewed in this article are nonlinear methods that, unlike linear methods such as Fourier, wavelets, and dictionary models, explicitly model signal intensities and their locations. Furthermore, they are often based on the theory of optimal mass transport from which fundamental principles can be put to use. Thus, they hold the promise to ultimately play a significant role in the development of a theoretical foundation for certain subclasses of modern learning and estimation problems.

A brief historical note

The optimal mass transport problem seeks the most efficient way of transforming one distribution of mass to another, relative to a given cost function. The problem was initially studied by the French mathematician Gaspard Monge in his seminal work "Mémoire sur la Théorie des Déblais et des Remblais" [35] in 1781. In 1942, Leonid V. Kantorovich, who, at that time, was unaware of Monge's work, proposed a general formulation of the problem by considering optimal mass transport plans, which, as opposed to Monge's formulation, allows for mass splitting [23]. Kantorovich shared the 1975 Nobel Prize in Economic Sciences with Tjalling Koopmans for his work in the optimal allocation of scarce resources. Kantorovich's contribution is considered "the birth of the modern formulation of optimal transport" [49], and it made the optimal mass transport problem an active field of research in the following years.

A significant portion of the theory of the optimal mass transport problem was developed in the 1990s, starting with Brenier's seminal work on the characterization, existence, and uniqueness of optimal transport maps [9], followed by Caffarelli's work on regularity conditions of such mappings [10] and Gangbo and McCann's work on a geometric interpretation of the problem [20]. A more thorough history and background on the optimal mass transport problem can be found in Villani's book *Optimal Transport: Old and New* [49] and Santambrogio's book *Optimal Transport for Applied Mathematicians* [45]. The significant contributions in mathematical foundations of the optimal transport problem together with recent advancements in numerical methods [6], [14], [31], [37] have spurred the recent development of numerous data-analysis techniques for modern estimation and detection (e.g., classification) problems.

Formulation of the problem and methodology

While reviewing both the continuous and discrete formulations of the optimal transport problem (i.e., Monge's and Kantorovich's formulations), the geometrical characteristics of the problem, and the transport-based signal/image embeddings, we have elected to avoid measure-theoretic notation, and other detailed mathematical language, in lieu of a more informal and intuitive description of the problem. However, it must be said that certain mathematical precision is required to best understand the differences between Monge's and Kantorivich's formulation, their geometric interpretations, and other points. The interested reader may find it useful to consult [24] for a more complete and mathematical description of the concepts explained in the following sections.

Optimal transport: Formulation

Over the past century or so, the theory of optimal transport (Earth mover's distance) has developed two main formulations, one utilizing a continuous map (Monge's formulation) and another utilizing what is called a *transport plan* (Kantarovich's formulation), for assigning the spatial correspondence necessary for the related transport problem. Although Monge's

continuous formulation is helpful in problems where a point-to-point assignment is desired, Kantarovich's formulation is more general and also covers the case of discrete (Dirac) masses (in our case, signal intensities). These not only differ in mathematical formulation but also have consequences with regard to their respective numerical solutions as well as applications.

Monge's continuous formulation

The Monge optimal mass transport problem is formulated as follows. Consider two signals or images I_0 and I_1 defined over their respective domains Ω_0 and Ω_1. Here, Ω_0 and Ω_1 are typically subsets of $\mathbb{R}^d$ and can often be taken as the unit square (or cube in three dimensions). Although a detailed measure-theoretic formulation is typically required (see [24]), we bypass the rigorous formulation here and simply assume that $I_0(x)$ and $I_1(y)$ correspond to signal intensities at positions $x \in \Omega_0$ and $y \in \Omega_1$. For digital signals, an interpolating model can be used to construct these functions defined over continuous domains from sampled discrete data. The signals are required to be nonnegative, i.e., $I_0(x) \geq 0 \;\; \forall x \in \Omega_0$ and $I_1(y) \geq 0 \;\; \forall y \in \Omega_1$. In addition, the total amount of signal (or mass) for both signals should be equal to the same constant (which is generally chosen to be 1): $\int_{\Omega_0} I_0(x)\,dx = \int_{\Omega_1} I_1(y)\,dy = 1$. In other words, I_0 and I_1 are assumed to be probability density functions (PDFs).

Monge's optimal transportation problem is to find a function $f: \Omega_0 \to \Omega_1$ that pushes I_0 onto I_1 and minimizes the objective function,

$$M(I_0, I_1) = \inf_{f \in MP} \int_{\Omega_0} c(x, f(x)) I_0(x)\,dx, \tag{1}$$

where $c: \Omega_0 \times \Omega_1 \to \mathbb{R}^+$ is the cost of moving pixel intensity $I_0(x)$ from x to $f(x)$ [Monge considered the Euclidean distance as the cost function in his original formulation, $c(x, f(x)) = |x - f(x)|$], and MP stands for a measure preserving map that moves all the signal intensity from I_0 to I_1. That is, for a subset $B \subset \Omega_1$ the MP requirement is that

$$\int_{\{x:\, f(x) \in B\}} I_0(x)\,dx = \int_B I_1(y)\,dy. \tag{2}$$

If f is one to one, this just means that for $A \subset \Omega_0$,

$$\int_A I_0(x)\,dx = \int_{f(A)} I_1(y)\,dy.$$

Such maps $f \in MP$ are sometimes called *transport maps* or *mass-preserving maps*. Simply put, the Monge formulation of the problem seeks to rearrange signal I_0 into signal I_1 while minimizing a specific cost function. In cases when f is smooth and one to one, then the requirement (2) can be written in a differential form as

$$\det(Df(x)) I_1(f(x)) = I_0(x) \tag{3}$$

almost everywhere, where Df is the Jacobian of f [see Figure 1(a)]. Note that both the objective function and the constraint in (1) are nonlinear with respect to $f(x)$. Hence, for more than a century, the answers to questions regarding existence and characterization of the Monge's problem remained unknown.

For certain measures, the Monge's formulation of the optimal transport problem is ill posed in the sense that there is no transport map to rearrange one PDF to another. For instance, consider the case where I_0 is a Dirac mass and I_1 is not. Kantorovich's formulation alleviates this problem by finding the optimal transport plan as opposed to the transport map.

Kantorovich's formulation

Kantorovich formulated the transport problem by optimizing over transportation plans, which we denote as γ. One can think of γ as the joint distribution of I_0 and I_1 describing how much mass is being moved to different coordinates; i.e., let A be a subset of Ω_0 and similarly $B \subseteq \Omega_1$. For notational simplicity, we will not make a distinction between a probability distribution and its density. More precisely, we associate a probability distribution to a signal I_0 by $I_0(A) = \int_A I_0(x)\,dx$.

The quantity $\gamma(A \times B)$ tells us how much mass in set A is being moved to set B. Here, the MP constraint can be expressed as $\gamma(\Omega_0 \times B) = I_1(B)$ and $\gamma(A \times \Omega_1) = I_0(A)$. Kantorovich's formulation for the optimal transport problem can then be written as

$$K(I_0, I_1) = \min_{\gamma \in MP} \int_{\Omega_0 \times \Omega_1} c(x, y)\,d\gamma(x, y). \tag{4}$$

Note that the integration notation $d\gamma(x, y)$ is meant to represent the fact that this integral is more general than the routine

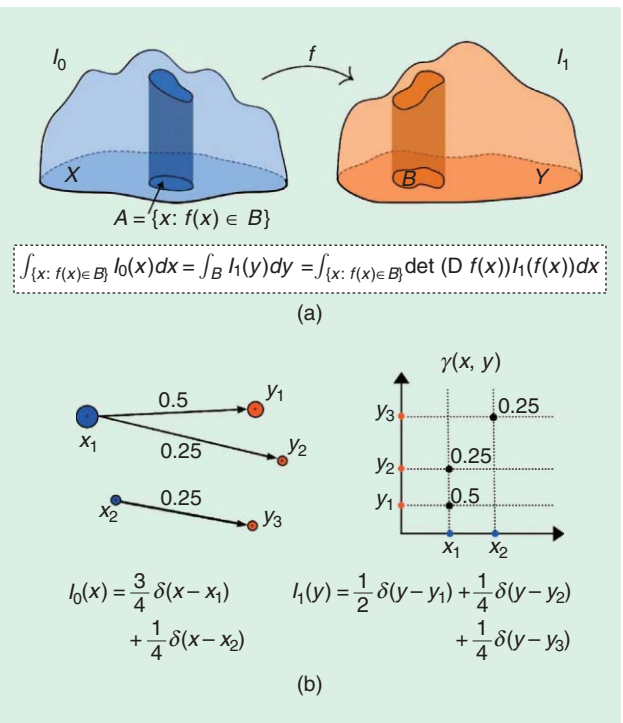


FIGURE 1. (a) The Monge transport map and (b) Kantorovich's transport plan.

Riemman-type integral commonly used in signal processing, and the integral can cover integration over domains that are more general. The minimizer of the optimization problem above, γ^*, is called the *optimal transport plan*. However, unlike the Monge problem, in Kantorovich's formulation, the objective function and the constraints are linear with respect to $\gamma(x,y)$. Moreover, Kantorovich's formulation is in the form of a convex optimization problem. We also note that the Monge problem is more restrictive than the Kantorovich problem; i.e., in Monge's version, mass from a single location in Ω_0 is being sent to a single location in Ω_1. Kantorovich's formulation, however, considers transport plans that can deal with arbitrary measurable sets and has the ability to distribute mass from the one location in one density to multiple locations in another [see Figure 1(b)]. For any transport map $f: \Omega_0 \to \Omega_1$ there is an associated transport plan, determined by

$$\gamma(A \times B) = \int_{\{x \in A:\, f(x) \in B\}} I_0(x)\,dx. \tag{5}$$

Furthermore, when an optimal transport map f^* exists, it can be shown that the transport plan γ^* derived from (5) is an optimal transportation plan [49].

The Kantorovich problem is especially interesting in a discrete setting, i.e., for PDFs of the form $I_0 = \sum_{i=1}^{M} p_i \delta(x - x_i)$ and $I_1 = \sum_{j=1}^{N} q_j \delta(y - y_j)$, where $\delta(x)$ is the Dirac delta function. Generally speaking, for such PDFs a transport map that pushes I_0 into I_1 does not exist. In these cases, mass splitting, as allowed by the Kantorovich formulation, is necessary [see Figure 1(b)]. The Kantorovich problem can be written as

$$\begin{aligned} K(I_0, I_1) &= \min_{\gamma} \sum_i \sum_j c(x_i, y_j)\gamma_{ij} \\ \text{s.t.} \sum_j \gamma_{ij} &= p_i, \ \sum_i \gamma_{ij} = q_j \\ \gamma_{ij} &\geq 0, i = 1,\ldots,M, j = 1,\ldots,N, \end{aligned} \tag{6}$$

where γ_{ij} identifies how much of the mass particle m_i at x_i needs to be moved to y_j [see Figure 1(b)]. The optimization above has a linear objective function and linear constraints; therefore, it is a linear programming problem. This problem is convex (which, in practice, translates to a relatively easier process of finding a global minimum), but not strictly so, and the constraint provides a polyhedral set of $M \times N$ matrices. In practice, a nondiscrete measure is often approximated by a discrete measure, and the Kantorovich problem is solved through the linear programming optimization expressed in (6).

Basic properties

Consider a transportation cost $c(x, y)$ that is continuous and bounded from below. Given two signals I_0 and I_1 as previously shown, there always exists a transportation plan minimizing (4). This holds true for both when signals I_0 and I_1 are functions and when they are discrete probability distributions [49]. Another important question is regarding the existence of an optimal transport map instead of a plan. Brenier [9] addressed this problem for the special case where $c(x,y) = |x - y|^2$. Bernier's results were later relaxed to more general cases by Gangbo and McCann [20], which led to the following theorem.

Theorem

Let I_0 and I_1 be nonnegative functions of the same total mass and with bounded support. When $c(x,y) = h(x - y)$ for some strictly convex function h, then there exists a unique optimal transportation map f^* minimizing (1). In addition, the optimal transport plan is unique and given by (5). Moreover, if $c(x,y) = |x - y|^2$, then there exists a (unique up to adding a constant) convex function ϕ such that $f^* = \nabla\phi$. A proof is available in [20] and [49].

Optimal mass transport: Geometric properties

Wasserstein metric

Let Ω be a bounded subset of $\mathbb{R}^d$ on which the signals are defined. As an example, for signals ($d = 1$) or images ($d = 2$), this can simply be the space $[0,1]^d$. Let $P(\Omega)$ be the set of probability densities supported on Ω. The p-Wasserstein metric, Wp, for $p \geq 1$ on $P(\Omega)$ is then defined as using the optimal transportation problem (4) with the cost function $c(x,y) = |x - y|^p$. For I_0 and I_1 in $P(\Omega)$,

$$W_p(I_0, I_1) = \left(* \inf_{\gamma \in MP} \int_{\Omega \times \Omega} |x - y|^p \, d\gamma(x, y) \right)^{\frac{1}{p}}.$$

For any $p \geq 1$, Wp is a metric on $P(\Omega)$. The metric space $(P(\Omega), W_p)$ is referred to as the p-Wasserstein space. To understand the nature of the optimal transportation distances, it is useful to note that for any $p \geq 1$, the convergence with respect to W_p is equivalent to the weak convergence of measures; i.e., $W_p(I_n, I) \to 0$ as $n \to \infty$ if and only if for every bounded and continuous function $f: \Omega \to \mathbb{R}$

$$\int_\Omega f(x) I_n(x)\,dx \to \int_\Omega f(x) I(x)\,dx.$$

For the specific case of $p = 1$, the p-Wasserstein metric is also known as the *Monge–Rubinstein metric* [49] or the *Earth mover's distance* [44]. The p-Wasserstein metric in one dimension has a simple characterization. For one-dimensional (1-D) signals I_0 and I_1, the optimal transport map has a closed-form solution. Let F_i be the cumulative distribution function of I_i for $i = 0, 1$, i.e.,

$$F_i(x) = \int_{\inf(\Omega)}^{x} I_i(x)\,dx \quad \textit{for } i = 0, 1.$$

Note that this is a nondecreasing function going from 0 to 1. We define the pseudoinverse of F_0 as follows: for $z \in (0,1)$, $F^{-1}(z)$ is the smallest x for which $F_0(x) \geq z$, i.e.,

$$F_0^{-1}(z) = \inf\{x \in \Omega : F_0(x) \geq z\}.$$

If $I_0 > 0$, then F_0 is continuous and increasing (and thus invertible), and the inverse of the function F_0 is equal to

the pseudoinverse we just defined. In other words, the pseudoinverse is a generalization of the notion of the inverse of a function. The pseudoinverse (i.e., the inverse if $I_0 > 0$ and $I_1 > 0$) provides a closed-form solution for the p-Wasserstein distance:

$$W_p(I_0, I_1) = \left(\int_0^1 |F_0^{-1}(z) - F_1^{-1}(z)|^p dz\right)^{\frac{1}{p}}. \tag{7}$$

The closed-form solution of the p-Wasserstein distance in one dimension is an attractive property, as it alleviates the need for optimization. This property was employed in the sliced-Wasserstein metrics as defined below.

Sliced-Wasserstein metric

The idea behind the sliced-Wasserstein metric is to first obtain a set of 1-D representations for a higher-dimensional probability distribution through projections (slicing the measure) and then calculate the distance between two input distributions as a functional on the Wasserstein distance of their 1-D representations. In this sense, the distance is obtained by solving several 1-D optimal transport problems, which have closed-form solutions.

The projection of high-dimensional PDFs is closely related to the well-known Radon transform in the imaging and image processing community [8], [25]. The d-dimensional Radon transform $\mathcal{R}$ maps a function $I \in L_1(\mathbb{R}^d)$ where $L_1(\mathbb{R}^d) := \{I: \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} |I(x)|\, dx \le \infty\}$ into the set of its integrals over the hyperplanes of $\mathbb{R}^n$. It is defined as

$$\mathcal{R}I(t,\theta) := \int_{\mathbb{R}} I(t\theta + s\theta^{\perp})\, ds, \quad \forall t \in \mathbb{R}, \forall \theta \in \mathbb{S}^{d-1};$$

here, $\theta^{\perp}$ is the subspace orthogonal to θ, and $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Note that $\mathcal{R}: L_1(\mathbb{R}^d) \to L_1(\mathbb{R} \times \mathbb{S}^{d-1})$. In other words, the Radon transform projects a PDF, $I \in P(\mathbb{R}^d)$, where $d > 1$, into an infinite set of 1-D PDFs $\mathcal{R}I(.,\theta)$. The sliced-Wasserstein metric for PDFs I_0 and I_1 on $\mathbb{R}^d$ is then defined as

$$SW_p(I_0, I_1) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}I_0(.,\theta), \mathcal{R}I_1(.,\theta))\, d\theta\right)^{\frac{1}{p}},$$

where $p \ge 1$, and W_p is the p-Wasserstein metric, which, for 1-D PDFs, $\mathcal{R}I_0(.,\theta)$ and $\mathcal{R}I_1(.,\theta)$ has a closed-form solution [see (7)]. For more details and definitions of the sliced-Wasserstein metric, we refer the reader to [8], [25] and [29].

Wasserstein spaces, geodesics, and Riemannian structure

In this section, we assume that Ω is convex. Here, we highlight that the p-Wasserstein space $(P(\Omega), W_p)$ is not just a metric space but has additional geometric structure. In particular, for any $p \ge 1$ and any $I_0, I_1 \in P(\Omega)$, there exists a continuous path (interpolation) between I_0 and I_1 whose length is the distance between I_0 and I_1.

Furthermore, the space with $p = 2$ is special because it possesses a structure of a formal, infinite dimensional, Riemannian manifold. That structure was first noted by Otto [38], who developed the formal calculations for using this structure. The precise description of the manifold of probability measures endowed with Wasserstein metric can be found in [1].

Next, we review the two main notions that have a wide use. We characterize the geodesics in $(P(\Omega), W_p)$, and in the case of $p = 2$, we describe what is the local, Riemannian metric of $(P(\Omega), W_2)$. Finally, we state the seminal result of Benamou and Brenier [5], who provided a characterization of geodesics via action minimization, which is useful in computations and also gives an intuitive explanation of the Wasserstein metric.

We first recall the definition of the length of a curve in a metric space. Let (X, d) be a metric space and $I: [a,b] \to X$. Then the length of I, denoted by $L(I)$ is

$$L(I) = \sup_{m \in \mathbb{N}, a = t_0 < t_1 < \cdots < t_m = b} \sum_{i=1}^{m} d(I(t_{i-1}), I(t_i)).$$

A metric space (X, d) is a geodesic space if, for any I_0 and I_1, there exists a curve $I: [0,1] \to X$ such that $I(0) = I_0, I(1) = I_1$ and for all $0 \le s < t \le 1, d(I(s), I(t)) = L(I|_{[s,t]})$. In particular, the length of I is equal to the distance from I_0 to I_1. Such a curve I is called a *geodesic*. The existence of geodesics is useful because it allows one to define the average of I_0 and I_1 as the midpoint of the geodesic connection between them.

An important property of $(P(\Omega), W_p)$ is that it is a geodesic space and that geodesics are easy to characterize. Specifically, they are given by the displacement interpolation (also known as a *McCann interpolation*). When a unique transportation map f^* from I_0 to I_1 exists that minimizes (1) for $c(x,y) = |x - y|^p$, the geodesic is obtained by moving the mass at constant speed from x to $f^*(x)$. More precisely, for $t \in [0,1]$ and $x \in \Omega$ let

$$f_t^*(x) = (1-t)x + tf^*(x)$$

be the position at time t of the mass initially at x. Note that f_0^* is identity mapping and $f_1^* = f^*$. Pushing forward the mass by f_t^*, which by (3) has the form

$$I_t(f_t^*(x)) = \frac{I_0(x)}{\det(Df_t^*(x))}$$

if f^* is smooth, provides the desired geodesic from I_0 to I_1. The velocity of each particle $\partial_t f_t^* = f^*(x) - x$ is the displacement of the optimal transportation map. Figure 2 conceptualizes the geodesic between two PDFs in $P(\Omega)$ and visualizes it for three different pairs of PDFs.

An important fact regarding the 2-Wasserstein space is Otto's presentation of a formal Riemannian metric for this space [38]. It involves shifting to a Lagrangian point of view. To explain, consider the path $I(x, t)$ in $P(\Omega)$ with $I(x, t)$ smooth. Then $s(x,t) = \partial I/\partial t(x,t)$ can be considered a tangent vector to the manifold or a density perturbation. Instead of thinking

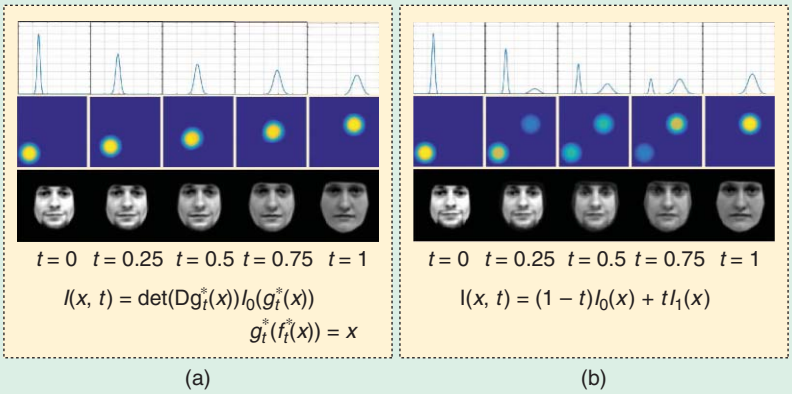


FIGURE 2. Geodesics in (a) the 2-Wasserstein space and in (b) the Euclidean space between various 1-D and two-dimensional (2-D) PDFs. Note that the geodesic in the 2-Wasserstein space captures the nonlinear structure of the signals and images and provides a natural morphing. (Face portraits courtesy of the public CMU Pose, Illumination, and Expression database.)

of increasing/decreasing the density, this perturbation can be viewed as resulting from moving the mass by a vector field. In other words, consider vector fields $v(x, t)$ such that

$$s = -\nabla\cdot(Iv). \tag{8}$$

There are many such vector fields. Otto defined the size of $s(\cdot,t)$ as the square root of the minimal kinetic energy of the vector field that produces the perturbation to density s, i.e.,

$$\langle s,s\rangle = \min_{v\text{satisfies}(8)} \int |v|^2 I dx. \tag{9}$$

Utilizing the Riemmanian manifold structure of $P(\Omega)$ together with the inner product presented in (9), the 2-Wasserstein metric can be reformulated into finding the minimizer of the following action among all curves in $P(\Omega)$ connecting I_0 and I_1 [5],

$$W_2^2(I_0,I_1) = \inf_{I,v} \int_0^1 \int_\Omega I(x,t)|v(x,t)|^2 dxdt$$
$$\text{s.t.}\quad \partial_t I + \nabla\cdot(Iv) = 0$$
$$I(\cdot,0) = I_0(\cdot), I(\cdot,1) = I_1(\cdot),$$

where the first constraint is the well-known continuity equation.

Optimal transport: Embeddings and transforms

The optimal transport problem and, specifically, the 2-Wasserstein metric and the sliced-2-Wasserstein metric have been recently used to define nonlinear transforms for signals and images [25], [27], [40], [50]. In contrast to commonly used linear signal transformation frameworks (e.g., Fourier and wavelet transforms) that employ signal intensities only at fixed coordinate points, thus adopting an Eulerian point of view, the idea behind transport-based transforms is to consider the intensity variations together with the locations of the intensity variations in the signal. Therefore, such transforms adopt a Lagrangian point of view for analyzing signals; i.e., they are able to move signal (pixel) intensities around. Moreover, the transforms can be viewed as Euclidean embeddings for the data, under the previously described transport-related metric space structure. The benefit of such a Euclidean embedding is that they facilitate the application of many standard data-analysis algorithms (e.g., learning). Here, we briefly describe these transforms and some of their prominent properties.

The linear optimal transportation framework

The linear optimal transportation (LOT) framework was proposed by Wang et al. [50]. The framework was used in [4] and [39] for pattern recognition in biomedical images and specifically histopathology and cytology images. Later, it was extended in [27] as a generic framework for pattern recognition, and it was used in [26] for the single-frame superresolution reconstruction of face images. The LOT framework, which provides an invertible Lagrangian transform for images, was initially proposed as a method to simultaneously amend the computationally expensive requirement of calculating pairwise 2-Wasserstein distance between N signals for pattern recognition purposes and to allow for the construction of generative models for images involving textures and shapes. For a given set of images $I_i \in P_2(\Omega)$, for $i = 1,...,N$, and a fixed template I_0, all nonnegative and having been normalized to have the same sum, the transform projects the images to the tangent space at I_0. The projections are acquired by finding the optimal velocity fields corresponding to the optimal transport plans between I_0 and each image in the set.

The framework provides a linear embedding for $P_2(\Omega)$ with respect to a fixed signal $I_0 \in P_2(\Omega)$. This means that the Euclidean distance between an embedded signal, denoted as $\tilde{I}_i$, and the fixed reference, I_0, is equal to $W_2(I_0, I_i)$, and the Euclidean distance between two embedded normalized signals is, generally speaking, an approximation of their 2-Wasserstein distance. The geometric interpretation of the LOT framework is presented in Figure 3. The linear embedding then facilitates the application of linear techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) to probability measures.

The cumulative distribution transform

Park et al. [40] considered the LOT framework for 1-D PDFs (positive signals normalized to integrate to 1), and since in dimension one the transport maps are explicit, they were able to characterize the properties of the transformed densities. Similar to the LOT framework, let I_i for $i = 1,...,N$ and I_0 be signals (PDFs) defined on $\mathbb{R}$. The framework first

calculates the optimal transport maps between I_i and I_0 using $f_i(x) = F_i^{-1} \circ F_0(x)$ for all $i = 1, ..., N$. Then the forward and inverse transport-based transform, denoted as the cumulative distribution transform (CDT) by Park et al. [40], for these density functions with respect to the fixed template I_0 is defined as

$$\begin{cases} \tilde{I}_i = (f_i - Id)\sqrt{I_0} & \text{(Analysis)} \\ I_i = (f_i^{-1})'(I_0 \circ f_i^{-1}) & \text{(Synthesis)'} \end{cases}$$

where $(I_0 \circ f_i^{-1})(x) = I_0(f_i^{-1}(x))$. Note that the L_2-norm (Euclidean distance) of the transformed signals, $\tilde{I}_i$, corresponds to the 2-Wasserstein distance between I_0 and I_i. In contrast to the higher-dimensional LOT, however, the Euclidean distance between two transformed (embedded) signals $\tilde{I}_i$ and $\tilde{I}_j$, is the exact 2-Wasserstein distance between I_i and I_j (see [40] for a proof) and not just an approximation. Hence, the transformation is isometric (preserves) with respect to the 2-Wasserstein metric. This isometric nature of the CDT was utilized in [28] to provide positive definite kernels for machine learning of n-dimensional signals.

From a signal processing point of view, the CDT is a nonlinear signal transformation that captures certain nonlinear variations in signals including translation and scaling. Specifically, it gives rise to the transformation pairs presented in Table 1. From Table 1, one can observe that although $I(t-\tau)$ is nonlinear in τ (when $I(.)$ is not a linear function), its CDT representation $\tilde{I}(t) + \tau\sqrt{I_0(t)}$ becomes affine in τ (a similar effect is observed for scaling). In effect, the Lagrangian transformations (compositions) in original signal space are rendered into Eulerian perturbations in transform space, borrowing from the partial differential equation (PDE) parlance. Furthermore, Park et al. [40] demonstrated that the CDT facilitates certain pattern recognition problems. More precisely, the transformation turns certain not linearly separable and disjoint classes of signals into linearly separable ones. Formally, let C be a set of 1-D maps, and let $P, Q \subset P_2(\Omega)$ be sets of positive PDFs born from two positive PDFs $p_0, q_0 \in P_2(\Omega)$ (which we denote as mother density functions or signals) as

$$P = \{p \mid p = h'(p_0 \circ h), \forall h \in C\},$$
$$Q = \{q \mid q = h'(q_0 \circ h), \forall h \in C\}.$$

If there exists no $h \in C$ for which $p_0 = h'(q_0 \circ h)$, then the sets P and Q are disjoint but not necessarily linearly separable in the signal space. A main result of [40] states that the signal classes P and Q are guaranteed to be linearly separable in the transform space (regardless of the choice of the reference signal I_0) if C satisfies the following conditions:

1) $h \in C \Rightarrow h^{-1} \in C$
2) $h_1, h_2 \in C \Rightarrow \rho h_1 + (1-\rho)h_2 \in C, \ \forall \rho \in [0,1]$
3) $h_1, h_2 \in C \Rightarrow h_1(h_2), h_2(h_1) \in C$
4) $h'(p_0 \circ h) \neq q_0, \ \forall h \in C.$

The set of translations $C = \{f \mid f(x) = x + \tau, \tau \in \mathbb{R}\}$ and scaling $C = \{f \mid f(x) = ax, a \in \mathbb{R}^+\}$, for instance, satisfy the above conditions. We refer the reader to [40] for further information. Figure 4(a) and (b) demonstrates the linear separation property of the CDT. The signal classes P and Q are chosen to be the set of all translations of a single Gaussian and a Gaussian mixture including two Gaussian functions with a fixed mean difference, respectively. The discriminant subspace is calculated for these classes, and it is shown that although the signal classes are not linearly separable in the signal domain, they become linearly separable in the transform domain.

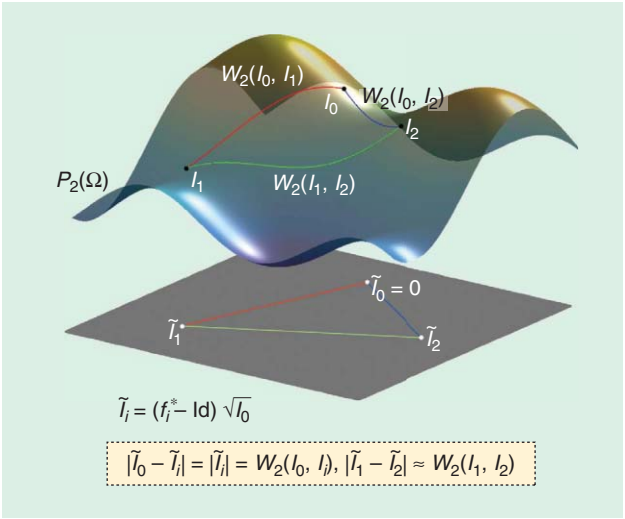


FIGURE 3. A graphical representation of the LOT framework. The framework embeds the PDFs (i.e., signals or images) I_i in the tangent space (i.e., the set of all tangent vectors) of $P(\Omega)$ with respect to a fixed PDF I_0. As a consequence, the Euclidean distance between the embedded functions $\tilde{I}_1$ and $\tilde{I}_2$ provides an approximation for the 2-Wasserstein distance, $W_2(I_1, I_2)$.

The Radon CDT

The CDT framework was extended to 2-D density functions (images) through the sliced-Wasserstein distance in [25] and was denoted the Radon CDT. It is shown in [25] that similar characteristics of the CDT, including the linear separation property, also hold for the Radon CDT. Figure 4 clarifies the linear separation property of the Radon CDT and demonstrate the capability of such transformations. In particular, Figure 4(c) and (d) shows a facial expression data set with two classes (i.e., neutral and smiling expressions) and its corresponding representations in the LDA discriminant subspace calculated from the images [Figure 4(c)] and the Radon CDT of the data set

Table 1. The CDT pairs. Note that the composition holds for all strictly monotonically increasing functions g.

Property	Signal Domain $I(x)$	CDT Domain $\tilde{I}(x)$
Translation	$I(x-\tau)$	$\tilde{I}(x) + \tau\sqrt{I_0(x)}$
Scaling	$aI(ax)$	$\frac{\tilde{I}(x)}{a} - x\frac{(a-1)}{a}\sqrt{I_0(x)}$
Composition	$g'(x)I(g(x))$	$(g^{-1}(\frac{\tilde{I}(x)}{\sqrt{I_0(x)}} + x) - x)\sqrt{I_0(x)}$

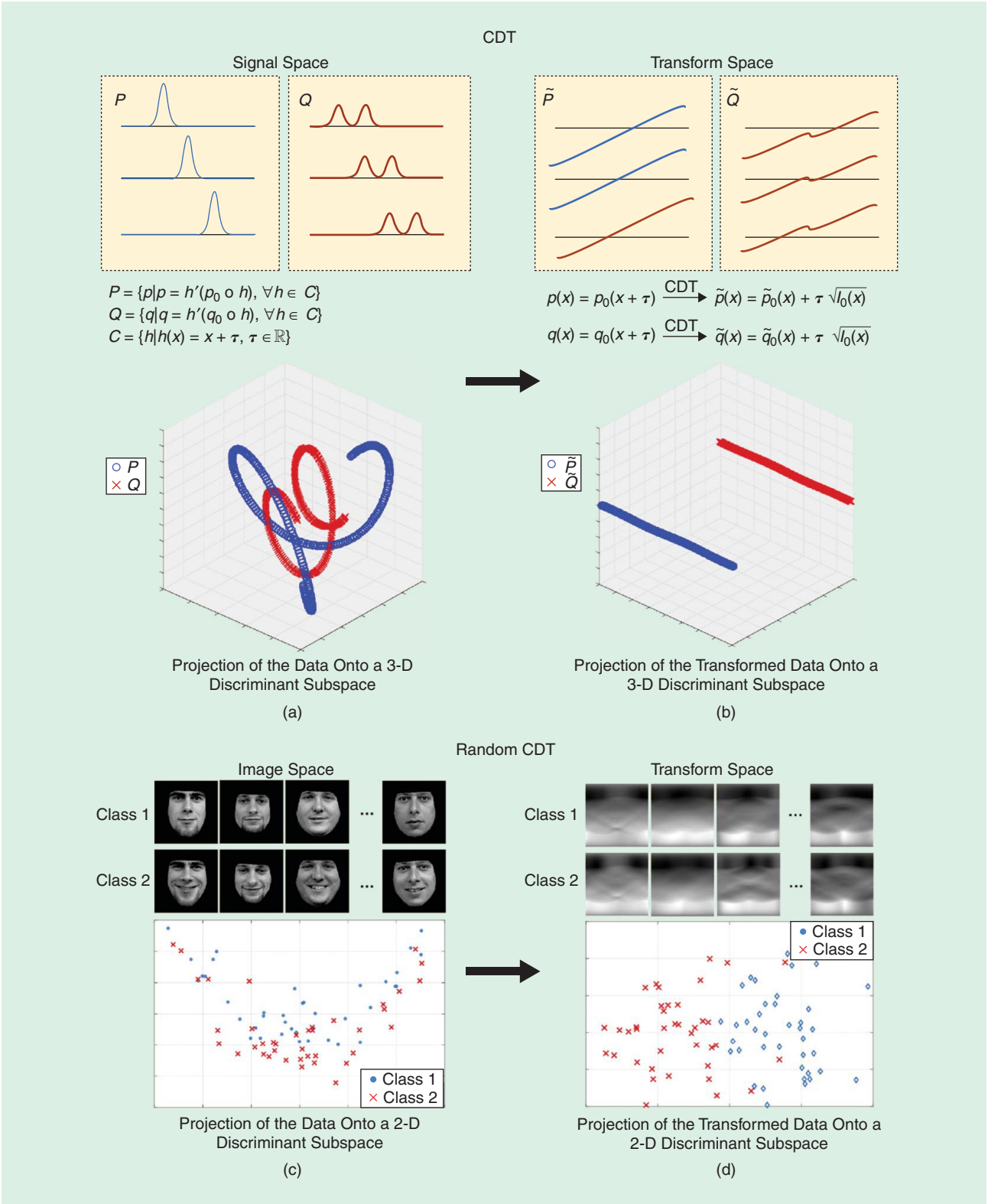


FIGURE 4. Examples for the linear separability characteristic of the CDT and the Radon CDT. The discriminant subspace for each case is calculated using the penalized-linear discriminant analysis. It can be seen that the nonlinear structure of the data is well captured in the transform spaces. (a) and (b) The linear separation property of the CDT. (c) A facial expression data set with two classes and its corresponding representations in the LDA discriminant subspace and (d) the Radon CDT of the data set and the corresponding representation of the transformed data in the LDA discriminant subspace. 3-D: three-dimensional. (Face portraits courtesy of the public CMU Pose, Illumination, and Expression database.)

and the corresponding representation of the transformed data in the LDA discriminant subspace [Figure 4(d)]. It is clear that the image classes become more linearly separable in the transform space. In addition, the cumulative percentage variation (CPV) of the data set in the image space, the Radon transform space, the Ridgelet transform space, and the Radon-CDT space are shown in Figure 5. The figure shows that the variations in the data set could be explained with fewer components in the Radon-CDT space.

Numerical methods

The development of robust and efficient numerical methods for computing transport-related maps, plans, metrics, and geodesics is crucial for the development of algorithms that can be used in practical applications. We next present several notable approaches for finding transportation maps and plans. Table 2 provides a high-level overview of these methods.

A linear programming problem

The linear programming problem is an optimization problem with a linear objective function and linear equality and inequality constraints. Several numerical methods exist for solving linear programming problems, among which are the simplex method and its variations and the interior-point methods. The computational complexity of the mentioned numerical methods, however, scales at best cubically in the size of the domain. Hence, assuming the measures considered have N particles, the number of unknowns γ_{ij}s is N^2 and the computational complexities of the solvers are at best $O(N^3 \log N)$ [14], [44]. The computational complexity of the linear programming methods is a very important limiting factor for the applications of the Kantorovich problem.

We note that, in the special case where I_0 and I_1 both have N equidistributed particles, the optimal transport problem simplifies to a one-to-one assignment problem that can be solved in $O(N^2 \log N)$. In addition, several multiscale approaches and sparse approximation approaches have recently been introduced to improve the computational performance of the linear programming solvers [37], [46].

Entropy-regularized solution

Cuturi's work [14] provides a fast and easy-to-implement variation of the Kantorovich problem by considering the transportation problem from a maximum-entropy perspective. The idea is to regularize the Wasserstein metric by the entropy of the transport plan. This modification simplifies the problem and enables much faster numerical schemes with complexity

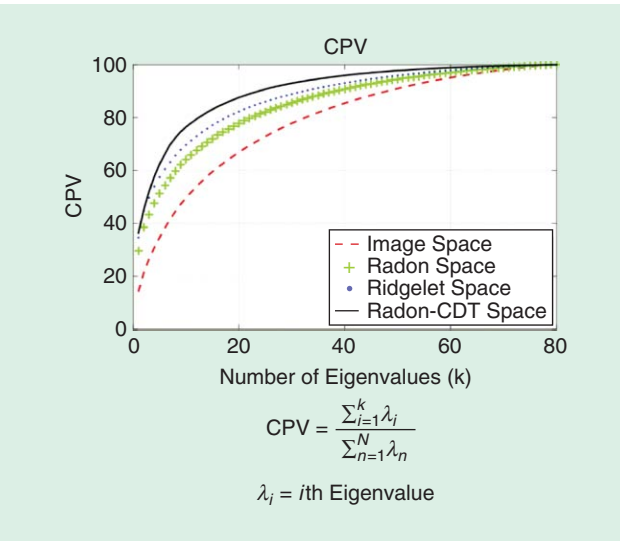


FIGURE 5. The cumulative percentage of the face data set in Figure 4 in the image space, the Radon transform space, the Ridgelet transform space, and the Radon-CDT transform space.

Table 2. The key properties of various numerical approaches.

Comparison of Numerical Approaches

Method	Remark
Linear programming	Applicable to general costs. Good approach if the PDFs are supported at very few sites.
Multiscale linear programming	Applicable to general costs. Fast and robust method, though truncation involved can lead to imprecise distances.
Auction algorithm	Applicable only when the number of particles in the source and the target is equal and all of their masses are the same.
Entropy-regularized linear programming	Applicable to general costs. Simple and performs very well in practice for moderately large problems. Difficult to obtain high accuracy.
Fluid mechanics	This approach can be adapted to generalizations of the quadratic cost, based on action along paths.
AHT minimization	Quadratic cost. Requires some smoothness and positivity of densities. Convergence is guaranteed only for infinitesimal step size.
Gradient descent on the dual problem	Quadratic cost. Convergence depends on the smoothness of the densities, hence a multiscale approach is needed for nonsmooth densities (i.e., normalized images).
Monge–Ampère solver	Quadratic cost. One in [7] is proved to be convergent. Accuracy is an issue due to the wide stencil used.
Semidiscrete approximation	An efficient way to find the map between a continuous and discrete signal [31].

AHT: Angenent, Haker, and Tannenbaum.

$O(N^2)$ [14] or $O(N\log(N))$ using the convolutional Wasserstein distance presented in [47] (compared to $O(N^3)$ of the linear programming methods), where N is the number of delta masses in each of the measures. The disadvantage is that it is difficult to obtain high-accuracy approximations of the optimal transport plan. The entropy-regularized p-Wasserstein distance, also known as the *Sinkhorn distance*, between PDFs I_0 and I_1 defined on the metric space (Ω, d) is defined as

$$W^p_{p,\lambda}(I_0, I_1) = \inf_{\gamma \in MP} \int_{\Omega\times\Omega} d^p(x, y)\gamma(x, y)\,dxdy + \lambda \int_{\Omega\times\Omega} \gamma(x, y)\ln(\gamma(x,y))\,dxdy, \quad (10)$$

where the regularizer is the negative entropy of the plan. We note that this is not a true metric since $W^p_{p,\lambda}(I_0,I_1) > 0$. Since the entropy term is strictly concave, the overall optimization in (10) becomes strictly convex. It is shown in [14] that the entropy-regularized p-Wasserstein distance in (10) can be reformulated as

$$W^p_{p,\lambda}(I_0,I_1) = \lambda * \inf_{\gamma\in MP} KL(\gamma \mid \mathcal{K}_\lambda),$$

where $\mathcal{K}_\lambda(x,y) = \exp(-d^p(x,y)/\lambda)$ and $KL(\gamma \mid \mathcal{K}_\lambda)$ is the Kullback–Leibler (KL) divergence between γ and $\mathcal{K}_\lambda$. In short, the regularizer enforces the plan to be within $1/\lambda$ radius in the KL-divergence sense from the transport plan $\gamma^*_\infty(x,y) = I_0(x)I_1(y)$.

Cuturi shows that the optimal transport plan γ in (10) is of the form $D_v\mathcal{K}_\lambda D_w$, where D_v and D_w are diagonal matrices with diagonal entries $v, w \in \mathbb{R}^N$ [14]; therefore, the number of unknowns in the regularized formulation is reduced from N^2 to $2N$. The new problem can then be solved through computationally efficient algorithms such as the iterative proportional fitting procedure, also known as the *iterative proportional fitting procedure algorithm*, or, alternatively, through the Sinkhorn–Knopp algorithm.

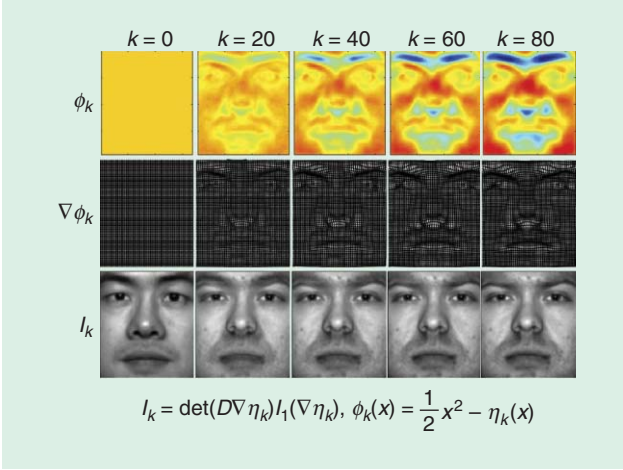


FIGURE 6. A visualization of the iterative update of the transport potential and correspondingly the transport displacement map through CWVB iterations. (Face portraits courtesy of the public Extended Yale Face Database B.)

Flow minimization (AHT)

Angenent, Haker, and Tannenbaum [2], proposed a flow minimization scheme to obtain the optimal transport map from the Monge problem. The method was used in several image-registration applications [22], pattern recognition [27], [50], and computer vision [26]. A brief review of the method is provided here.

Let $I_0: X \to \mathbb{R}^+$ and $I_1: Y \to \mathbb{R}^+$ be continuous probability densities defined on convex domains $X, Y \subseteq \mathbb{R}^d$. To find the optimal transport map, f^*, AHT starts with an initial transport map, $f_0: X \to Y$ calculated from the Knothe–Rosenblatt coupling [49]. Then it updates f_0 to minimize the transport cost while constraining it to remain a transport map from I_0 to I_1. The updated equation for finding the optimal transport map in AHT is calculated to be

$$f_{k+1}(x) = f_k(x) + \epsilon\frac{1}{I_0}Df_k(f_k - \nabla(\Delta^{-1}div(f_k))),$$

where ϵ is the step size, Df_k is the Jacobian matrix, and Δ^{-1} is the Poisson solver with Neumann boundary conditions. AHT show that for infinitesimal step size, ϵ, $f_k(x)$ converges to the optimal transport map. For a detailed derivation of the preceding equation, see [2] and [24].

The AHT method is, in essence, a gradient descent method on the Monge formulation of the optimal transport problem. Chartrand, Wohlberg, Vixie, and Bollt (CWVB) [11] proposed an alternative gradient-descent method based on Kantorovich's dual formulation of the transport problem that updates the optimal potential transport field, $\eta(x)$, where $f(x) = \nabla\eta(x)$. Figure 6 presents the iterations of the CWVB method for two face images taken from the YaleB face database.

Monge–Ampère equation

The Monge–Ampère PDE is defined as

$$\det(H\phi) = h(x, \phi, D\phi)$$

for some functional h and where $H\phi$ is the Hessian matrix of ϕ. The Monge–Ampère PDE is closely related to the Monge problem for the quadratic cost function. According to Bernier's theorem (discussed in the "Basic Properties" section), when I_0 and I_1 are absolutely continuous PDFs defined on sets $X, Y \subset \mathbb{R}^n$, the optimal transport map that minimizes the 2-Wasserstein metric is uniquely characterized as the gradient of a convex function $\phi: X \to Y$. Moreover, we showed that the mass-preserving constraint of the Monge problem can be written as $\det(Df)I_1(f) = I_0$. Combining these results, one can have

$$\det(D(\nabla\phi(x))) = \frac{I_0(x)}{I_1(\nabla\phi)}, \quad (11)$$

where $D\nabla\phi = H\phi$, and, therefore, the equation shown above is in the form of the Monge–Ampère PDE. Now, if ϕ is a convex function on X satisfying $\nabla\phi(X) = Y$ and solving (11), then $f^* = \nabla\phi$ is the optimal transportation map from I_0 to I_1.

The geometrical constraint on this problem is rather unusual in PDEs and is often referred to as the *optimal transport boundary conditions*. Several authors have proposed numerical methods to obtain the optimal transport map through solving the Monge–Ampère PDE in (11) [7], [33]. In particular, the scheme in [7] is monotone, has complexity $O(N)$ (up to logarithms), and is provably convergent. We conclude by remarking that several regularity results on the optimal transport maps were established through the Monge–Ampère equation (see [24] for references).

Semidiscrete approximation

Several works [31], [34] have considered the problem in which one PDF, I_0, has a continuous form while the other, I_1 is discrete, $I_1(y) = \sum q_i \delta(y - y_i)$. It turns out there exist weights w_i such that the optimal transport map $f: X \to Y$ can be described via a power diagram. More precisely, the set of x mapping to y_i is the following cell of the power diagram:

$$PD_w(y_i) = \{x: |x - y_i|^2 - w_i \le |x - y_j|^2 - w_j, \forall j\}.$$

The main observation is that the weights w_i are minimizers of the following unconstrained convex functional:

$$\sum_i \left(q_i w_i - \int_{PD_w(y_i)} (||x - y_i||^2 - w_i) I_0(x)\, dx \right).$$

Works by Mérigot [34] and Levy [31] use Newton-based schemes and multiscale approaches to minimize the functional. The need to integrate over the power diagram makes the implementation somewhat geometrically delicate. Nevertheless, a recent implementation by Levy [31] gives impressive results in terms of speed. This approach provides the transportation mapping (not just the approximation of a plan).

Applications

Image retrieval

One of the earliest applications of the optimal transport problem was in image retrieval. Rubner et al. [44] employed the discrete Wasserstein metric, which they denoted the Earth mover's distance, to measure the dissimilarity between image signatures. In image-retrieval applications, it is common practice first to extract features (i.e., color features, texture feature, shape features, and so on) and then generate high-dimensional histograms or signatures (histograms with dynamic/adaptive binning) to represent images. The retrieval task then simplifies to finding images with similar representations (e.g., small distance between their histograms/signatures). The Wasserstein metric is specifically suitable for such applications because it can compare histograms/signatures of different sizes (histograms with different binning). This unique capability turns the Wasserstein metric into an attractive candidate in image-retrieval applications [32], [44]. In [44], the Wasserstein metric was compared with common metrics such as Jeffrey's divergence, the χ^2 statistic, the L_1 distance, and the L_2 distance in an image-retrieval task, and it was shown that the Wasserstein metric achieves the highest precision/recall performance among all.

Speed of computation is an important practical consideration in image-retrieval applications. For almost a decade, the high computational cost of the optimal transport problem overshadowed its practicality in large-scale image-retrieval applications. Recent advancements in numerical methods, including the work of Merigot [34] and Cuturi [14], among many others, have reinvigorated optimal transport-based distances as a feasible and appealing candidate for large-scale image-retrieval problems.

Registration and morphing

Image registration deals with finding a common geometric reference frame between two or more images. It plays an important role in analyzing images obtained at different times or using different imaging modalities. Image registration and, more specifically, biomedical image registration are active areas of research. Registration methods find a transformation f that maximizes the similarity between two or more image representations (e.g., image intensities and image features). Among the plethora of registration methods, nonrigid registration methods are especially important given their numerous applications in biomedical problems. They can be used to quantify the morphology of different organs, correct for physiological motion, and allow for comparison of image intensities in a fixed coordinate space (atlas). Generally speaking, nonrigid registration is a nonconvex and nonsymmetric problem, with no guarantee of the existence of a globally optimal transformation.

Various works in the literature deploy the Monge problem for image warping and elastic registration. Utilizing the Monge problem in an image-warping/registration setting has a number of advantages. First, the existence and uniqueness of the global transformation (the optimal transport map) is known. Second, the problem is symmetric, meaning that the optimal transport map for warping I_0 to I_1 is the inverse of the optimal transport map for warping I_1 to I_0. Last, it provides a landmark-free and parameter-free registration scheme with a built-in mass preservation constraint. These advantages motivated several follow-up works to investigate the application of the Monge problem in image registration and warping [21], [22].

In addition to images, the optimal mass transport problem has also been used in point cloud and mesh registration [29] (see [24] for more references), which have various applications in shape analysis and graphics. In these applications, shape images (2-D or 3-D binary images) are first represented using either sets of weighted points (e.g., point clouds), using clustering techniques such as K-means or fuzzy C-means, or with meshes. Then a regularized variation of the optimal transport problem is solved to match such representations. The regularization on the transportation problem is often imposed to enforce the neighboring points (or vertices) to remain near each other after the transformation.

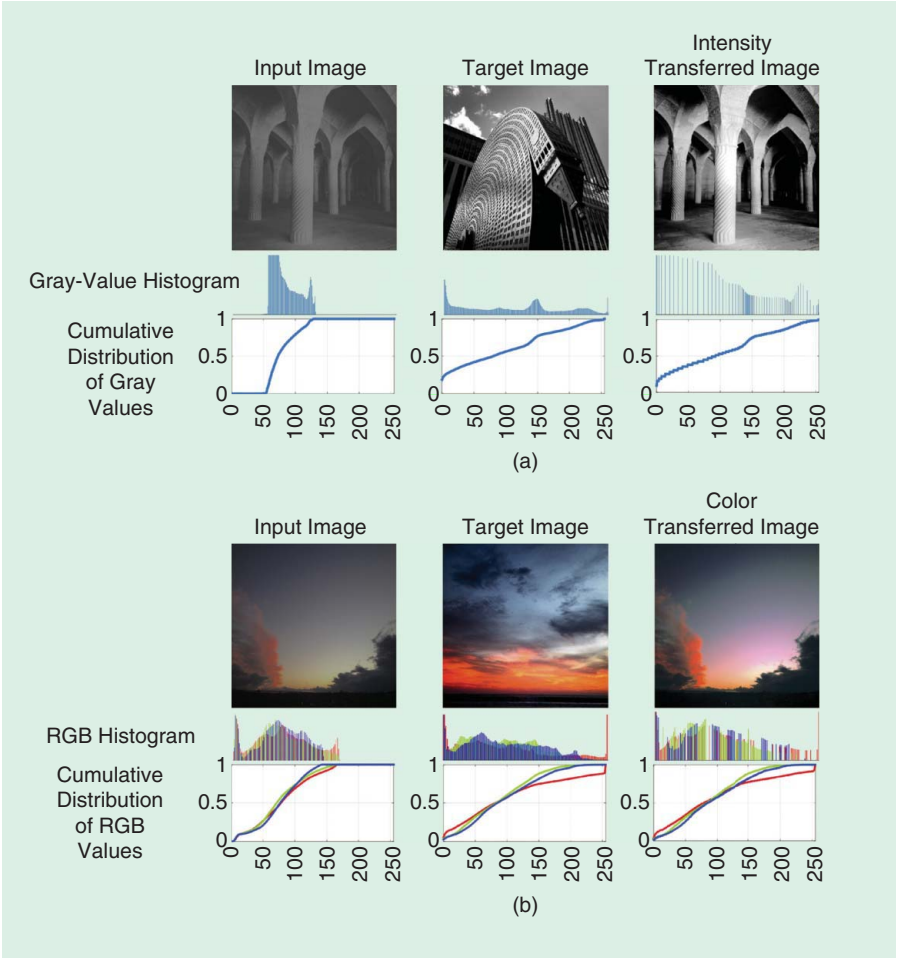


FIGURE 7. (a) Gray value and (b) color transfer via optimal transportation. RGB: red, green, blue.

Color transfer and texture synthesis

Texture mixing and color transfer are appealing applications of the optimal transport framework in image analysis, graphics, and computer vision. Here, we briefly discuss these applications.

Color transfer

The purpose of color transfer is to change the color palette of an image to impose the feel and look of another image. Color transfer is generally performed through finding a map, which morphs the color distribution of the first image into the second one. For grayscale images, the color-transfer problem simplifies to a histogram-matching problem, which is solved through the 1-D optimal transport formulation [16]. In fact, the classic problem of histogram equalization is a 1-D transport problem [16]. The color-transfer problem, on the other hand, is concerned with pushing the 3-D color distribution of the first image into the second one. This problem can also be formulated as an optimal transport problem, as demonstrated in [41] (see [24] for more references).

A complication that occurs in the color transfer on real images, however, is that a perfect match between color distributions of the images is often not satisfying, because a color-transfer map may not transfer the colors of neighboring pixels in a coherent manner and may introduce artifacts in the color-transferred image. Therefore, the color-transfer map is often regularized to make the transfer map spatially coherent [41]. Figure 7 shows a simple example of gray-value and color transfer via the optimal transport framework. It can be seen that the cumulative distribution of the gray-value and color-transferred images are similar to that of the input image.

Texture synthesis and mixing

Texture synthesis is the problem of synthesizing a texture image that is visually similar to an exemplar input-texture image and has various applications in computer graphics and image processing. Many methods have been proposed for texture synthesis, such as synthesis by recopy and synthesis by statistical modeling. Texture mixing, however, considers the problem of synthesizing a texture image from a collection of input-texture images in a way that the synthesized texture provides a meaningful integration of the colors and textures of the input-texture images. Metamorphosis is one of the successful approaches in texture mixing; it performs the mixing via identifying correspondences between elementary features (i.e., textons) among input textures and progressively morphing between the shapes of elements. In other approaches, texture images are first parametrized through a tight frame (often steerable wavelets), and statistical modeling is performed on the parameters.

Other successful approaches include random phase and spot noise texture modeling [18], which model textures as stationary Gaussian random fields. These models are based on the assumption that the visual texture perception is based on the spectral magnitude of the texture image. Therefore, utilizing the spectral magnitude of an input image and randomizing its phase will lead to a new synthetic texture image that is visually similar to the input image. Ferradans et al. [18] utilized this assumption together with the Wasserstein geodesics to interpolate between spectral magnitude of texture images and provide synthetic mixed texture images. Figure 8 shows an example of texture missing via the Wasserstein geodesic between the spectral magnitudes of the input-texture images. The in-between images are synthetically generated using the random-phase technique.

Image denoising and restoration

The optimal transport problem has also been used in several image-denoising and -restoration problems [30]. The goal in these applications is to restore or reconstruct an image from noisy or incomplete observation. Lellmann et al. [30] utilized the Kantorovich–Rubinsten discrepancy term together with a total variation (TV) term in the context of image denoising. They called their method *Kantorovich–Rubinstein-TV* (*KR-TV*) denoising. Note that the KR metric is closely related to the 1-Wasserstein metric (for 1-D signals they are equivalent). The KR term in their proposed functional provides a fidelity term for denoising, and the TV term enforces a piecewise constant reconstruction.

Transport-based morphometry

Given their suitability for comparing mass distributions, transport-based approaches for performing pattern recognition of morphometry encoded in image intensity values have also lately emerged. Recently described approaches for transport-based morphometry (TBM) [4], [27], [50] work by computing transport maps or plans between a set of images and a reference or template image. The transport plans/maps are then utilized as an invertible feature/transform onto which pattern recognition algorithms such as PCA or LDA can be applied. In effect, it utilizes the LOT framework described in the "The Linear Optimal Transportation Framework" section. These techniques have recently been employed to decode differences in cell and nuclear morphology for drug screening [4], cancer detection histopathology [39], and cytology images, as well as applications such as the analysis of galaxy morphologies [27].

Deformation-based methods have long been used in analyzing biomedical images. TBM, however, is different from those deformation-based methods in that it has numerically exact, uniquely defined solutions for the transport plans or maps used; i.e., images can be matched with little perceptible error. The same is not true in methods that rely on registration via the computation of deformations, given the significant topology differences commonly found in medical images. Moreover, TBM allows for comparison of the entire intensity information present in the images (shapes and textures), while deformation-based methods are usually employed to deal with shape differences. Figure 9 shows a schematic of the TBM steps applied to a cell nuclei data set. It can be seen that TBM is capable of modeling the variation in the data set. In addition, it enables one to visualize the classifier, which discriminates between image classes (in this case malignant versus benign).

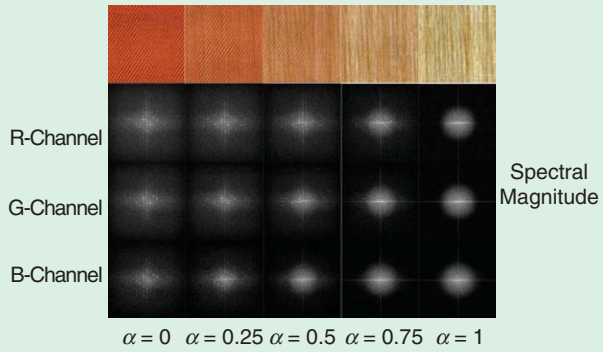


FIGURE 8. An example of texture mixing via optimal transport using the method presented in Ferradans et al. [18].

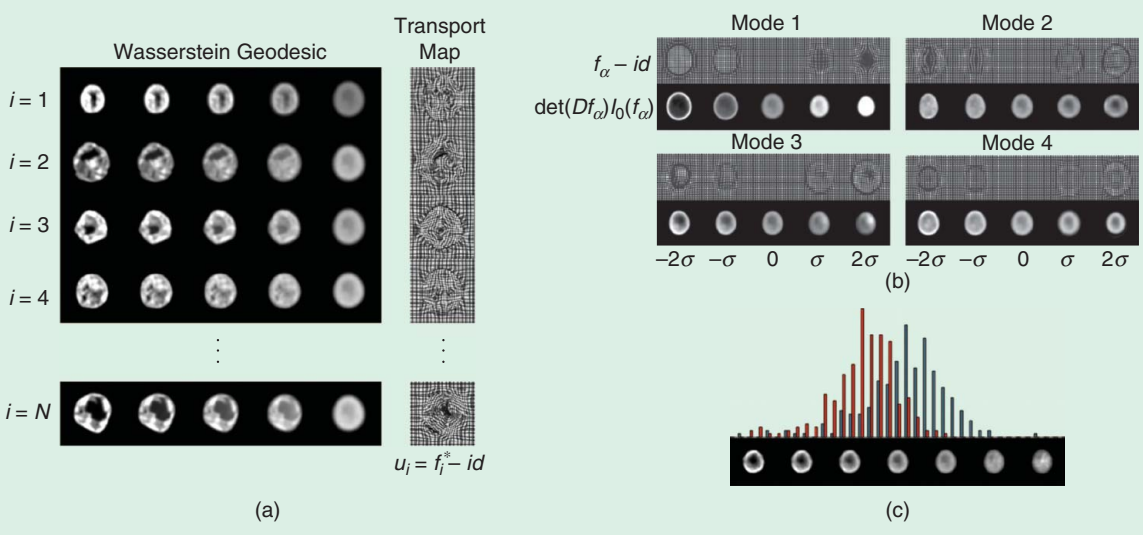


FIGURE 9. The schematic of the TBM framework. (a) The optimal transport maps between input images $I_1, \ldots, I_N$ and a template image I_0 are calculated. (b) and (c) Next, linear statistical modeling such as PCA, LDA, and canonical correlation analysis is performed on the optimal transport maps. The resulting transport maps obtained from the statistical modeling step are then applied to the template image to visualize the results of the analysis in the image space.

Superresolution

Superresolution is the process of reconstructing a high-resolution image from one or several corresponding low-resolution images. Superresolution algorithms can be broadly categorized into two major classes, multiframe superresolution and single-frame superresolution, based on the number of low-resolution images they require to reconstruct the corresponding high-resolution image. The TBM approach was used for single-frame superresolution in [26] to reconstruct high-resolution faces from very low-resolution-input face images. The authors utilized the TBM in combination with subspace learning techniques to learn a nonlinear model for the high-resolution face images in the training set.

In short, the method consists of a training and a testing phase. In the training phase, it uses high-resolution face images and morphs them to a template high-resolution face through optimal transport maps. Next, it learns a subspace for the calculated optimal transport maps. A transport map in this subspace can then be applied to the template image to synthesize a high-resolution face image. In the testing phase, the goal is to reconstruct a high-resolution image from the low-resolution input image. The method searches for a synthetic high-resolution face image (generated from the transport subspace) that provides a corresponding low-resolution image, which is similar to the input low-resolution image. Figure 10 shows the steps used in this method and demonstrates reconstruction results.

Machine learning and statistics

The optimal transport framework has recently attracted ample attention from the machine-learning and statistics communities [12], [19], [25], [28], [36]. Some applications of the optimal transport in these arenas include various transport-based learning methods [19], [28], [36], [48], domain adaptation, Bayesian

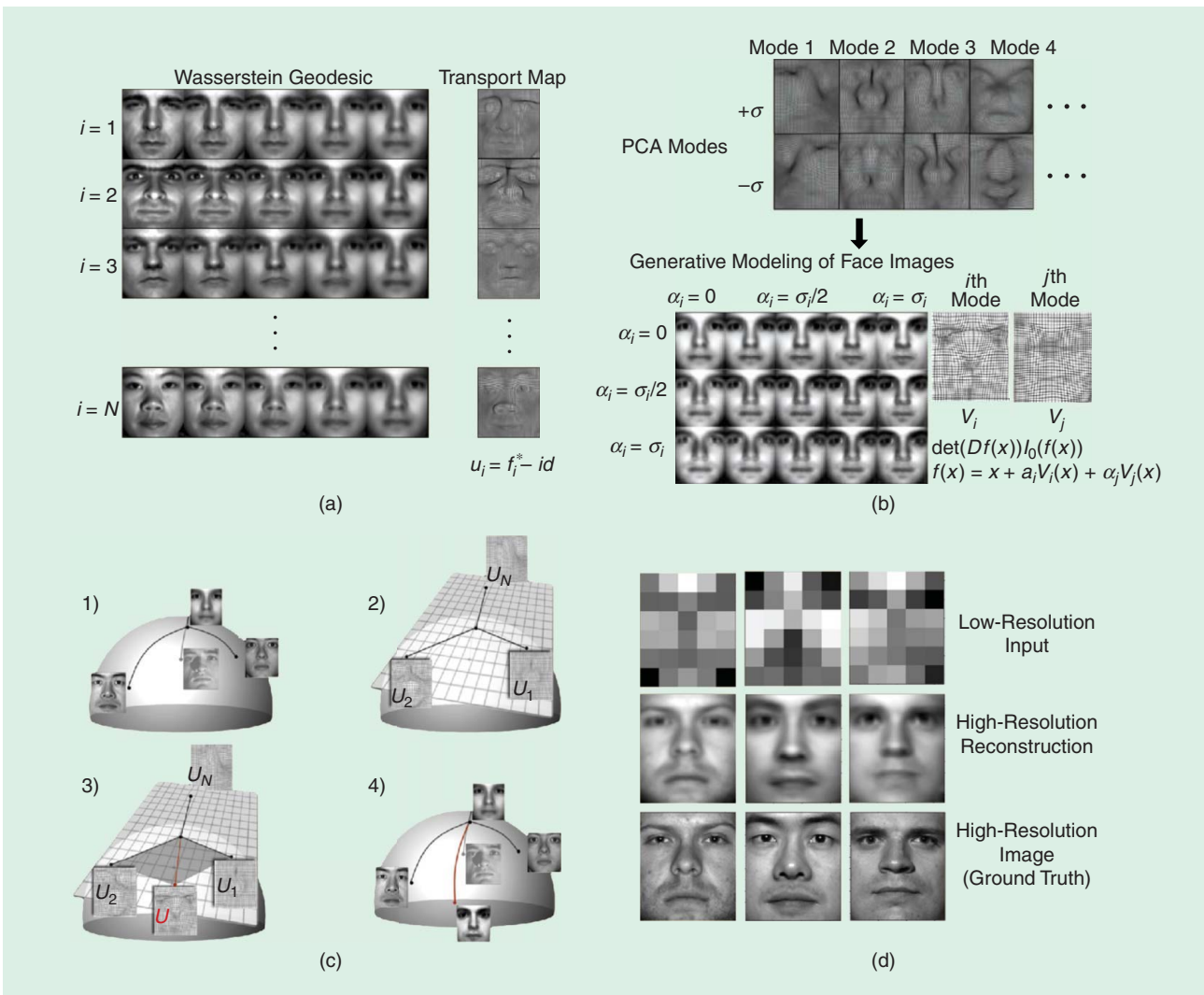


FIGURE 10. (a) In the training phase, optimal transport maps that morph the template image to high-resolution training face images are calculated. (b) Statistical modeling of transport maps. PCA is used to learn a linear subspace for transport maps for which a linear combination of obtained eigenmaps can be applied to the template image to obtain synthetic face images. (c) A geometric interpretation of the problem and (d) reconstruction results in transport-based single-frame superresolution. (Face portraits courtesy of the public Extended Yale Face Database B.)

inference [12], [13] and hypothesis testing [15], [42] among others. Here, we provide a brief overview of the recent developments of transport-based methods in machine learning and statistics.

Learning

Transport-based distances have recently been used in several works as a loss function for regression, classification, and other techniques. Montavon, Müller, and Cuturi [36], for instance, utilized the dual formulation of the entropy-regularized Wasserstein distance to train restricted Boltzmann machines (RBMs). Boltzmann machines are probabilistic graphical models (Markov random fields) that can be categorized as stochastic neural networks and are capable of extracting hierarchical features at multiple scales. RBMs are bipartite graphs that are special cases of Boltzmann machines, which define parameterized probability distributions over a set of d-binary input variables (observations) whose states are represented by h binary output variables (hidden variables). The parameters of RBMs are often learned through information theoretic divergences such as KL divergence. Montavon et al. [36] proposed an alternative approach through a scalable entropy-regularized Wasserstein distance estimator for RBMs and showed the practical advantages of this distance over the commonly used information divergence-based loss functions.

In another approach, Frogner et al. [19] used the entropy-regularized Wasserstein loss for multilabel classification. They proposed a relaxation of the transport problem to deal with unnormalized measures by replacing the equality constraints in (6) with soft penalties with respect to KL divergence. In addition, Frogner et al. [19] provided statistical bounds on the expected semantic distance between the prediction and the ground truth. In yet another approach, Kolouri et al. [28] utilized the sliced-Wasserstein metric and provided a family of positive definite kernels, denoted sliced-Wasserstein kernels, and showed the advantages of learning with such kernels. The sliced-Wasserstein kernels were shown to be effective in various machine-learning tasks, including classification, clustering, and regression.

Solomon et al. [48] considered the problem of graph-based semisupervised learning, in which graph nodes are partially labeled and the task is to propagate the labels throughout the nodes. Specifically, they considered a problem in which the labels are histograms. This problem arises, for example, in traffic density prediction, in which the traffic density is observed for a few stop lights over 24 h in a city and the city is interested in predicting the traffic density at the unobserved stop lights. They pose the problem as an optimization of a Dirichlet energy for distribution-valued maps based on the 2-Wasserstein distance and present a Wasserstein propagation scheme for semisupervised distribution propagation along graphs.

More recently, Arjovskly et al. [3] compared various distances, i.e., TV, KL divergence, Jenson–Shannon divergence, and the Wasserstein distance in training generative adversarial networks (GANs). They demonstrated (theoretically and numerically) that the Wasserstein distance leads to a superior performance compared to the later dissimilarity measures. They specifically showed that their proposed Wasserstein GAN does not suffer from common issues in such networks, including instability and mode collapse.

Domain adaptation

Domain adaptation is one of the fundamental problems in machine learning that has gained proper attention from the machine-learning research community in the past decade. Domain adaptation is the task of transferring knowledge from classifiers trained on available labeled data to unlabeled test domains with data distributions that differ from that of the training data. The optimal transport framework was recently presented as a potential major player in domain adaptation problems [12], [13]. Courty et al. [12], for instance, assumed that there exists a nonrigid transformation between the source and target distributions, and they find this transformation using an entropy-regularized optimal transport problem. They also proposed a label-aware version of the problem in which the transport plan is regularized so a given target point (testing exemplar) is associated only with source points (training exemplars) belonging to the same class. Courty et al. [12] showed that domain adaptation via regularized optimal transport outperforms the state-of-the-art results in several challenging domain adaptation problems.

Bayesian inference

Another interesting and emerging application of the optimal transport problem is in Bayesian inference [17]. In Bayesian inference, one critical step is the evaluation of expectations with respect to a posterior probability function, which leads to complex multidimensional integrals. These integrals are commonly solved through the Monte Carlo numerical integration, which requires independent sampling from the posterior distribution. In practice, sampling from a general posterior distribution might be difficult, so, therefore, the sampling is performed via a Markov chain that converges to the posterior probability after a certain number of steps. This leads to the celebrated Markov chain Monte Carlo (MCMC) method. The downside of the MCMC method is that the samples are not independent, and, hence, the convergence of the empirical expectation is slow. El Moselhy and Marzouk [17] proposed a transport-based method that evades the need for Markov-chain simulation by allowing direct sampling from the posterior distribution. The core idea in their work is to find a transport map (via a regularized Monge formulation) that pushes forward the prior measure to the posterior measure. Then, sampling the prior distribution and applying the transport map to the samples will lead to a sampling scheme from the posterior distribution. Figure 11 shows the basic idea behind these methods.

Hypothesis testing

The Wasserstein distance is used for goodness-of-fit testing in [15] and for two-sample testing in [42]. Ramdas et al. [42] presented connections between the entropy-regularized Wasserstein distance, multivariate Energy distance, and the kernel maximum mean discrepancy and provided a "distribution-free" univariate Wasserstein test statistic. These and other

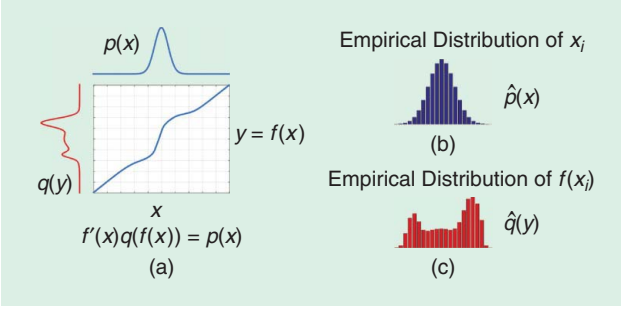


FIGURE 11. (a) The prior distribution p, the posterior distribution q, and the corresponding transport map f that pushes p into q. One million samples, x_i, were generated from distribution p. (b) The empirical distribution of these samples denoted as $\hat{p}$ and (c) the empirical distribution of transformed samples, $y_i = f(x_i)$, denoted as $\hat{q}$.

applications of transport-related concepts show the promise of the mathematical modeling technique in the design of statistical data-analysis methods to tackle modern learning problems. Finally, note that, in the interest of brevity, a number of other important applications of transport-related techniques were not discussed above but are certainly interesting in their own right. For a more detailed discussion and more references please refer to [24].

Summary and conclusions

Transport-related methods and applications have come a long way. Although earlier applications focused primarily in civil engineering and economics problems, they have recently begun to be employed in a wide variety of problems related to signal and image analysis and pattern recognition. In this article, seven main areas of application were reviewed: image retrieval, registration and morphing, color transfer and texture analysis, image restoration, TBM, image superresolution, and machine learning and statistics. Transport and related techniques have gained increased interest in recent years. Overall, researchers have found that the application of transport-related concepts can be helpful in solving problems in diverse applications. Given recent trends, it seems safe to expect that the number of application areas will continue to grow.

In its most general form, the transport-related techniques reviewed in this article can be thought as mathematical models for signals and images and in general data distributions. Transport-related metrics involve calculating differences not only of pixel or distribution intensities but also where they are located in the corresponding coordinate space (a pixel coordinate in an image or a particular axis in some arbitrary feature space). As such, the geometry (e.g., geodesics) induced by such metrics can give rise to dramatically different algorithms and data interpretation results. The interesting performance improvements recently obtained could motivate the search for a more rigorous mathematical understanding of transport-related metrics and applications.

The emergence of numerically precise and efficient ways of computing transport-related metrics and geodesics, as presented in the "Numerical Methods" section, also serves as an enabling mechanism. Coupled with the fact that several mathematical properties of transport-based metrics have been extensively studied, we believe that the foundation is set for their increased use as tools or building blocks based on which complex computational systems can be built. The confluence of these emerging ideas may spur a significant amount of innovation in a world where sensor and other data are becoming abundant and computational intelligence to analyze these is in high demand. We believe transport-based models will become an important component of the ever-expanding tool set available to modern signal-processing and data-science experts.

Acknowledgments

We gratefully acknowledge funding from the National Science Foundation (NSF) (CCF 1421502) and the National Institutes of Health (GM090033, CA188938) in contributing to a portion of this work. Dejan Slepčev also acknowledges funding by the NSF (DMS DMS-1516677).

Authors

Soheil Kolouri (skolouri@andrew.cmu.edu) received his B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2010 and his M.S. degree in electrical engineering in 2012 from Colorado State University, Fort Collins. He received his doctorate degree in biomedical engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2015, where his thesis, "Transport-Based Pattern Recognition and Image Modelling," won the Best Thesis Award. He is currently with HRL Laboratories, LLC, Malibu, California.

Se Rim Park (serimp@andrew.cmu.edu) received her B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011 and is currently a doctoral candidate in the Electrical and Computer Engineering Department at Carnegie Mellon University, Pittsburgh, Pennsylvania. She is mainly interested in signal processing and machine learning, especially designing new signal and image transforms and developing novel systems for pattern recognition.

Matthew Thorpe (M.Thorpe@warwick.ac.uk) received his B.S., M.S., and Ph.D. degrees in mathematics from the University of Warwick, United Kingdom, in 2009, 2012, and 2015, respectively, and his M.S.Tech. degree in mathematics from the University of New South Wales, Australia, in 2010. He is currently a postdoctoral associate within the Mathematics Department at Carnegie Mellon University, Pittsburgh, Pennsylvania.

Dejan Slepčev received his B.S. degree in mathematics from the University of Novi Sad, Serbia, in 1995, his M.A. degree in mathematics from the University of Wisconsin–Madison in 2000, and his Ph.D. degree in mathematics from the University of Texas at Austin in 2002. He is currently an associate professor in the Department of Mathematical Sciences at Carnegie Mellon University, Pittsburgh, Pennsylvania.

Gustavo K. Rohde (gustavo@virginia.edu) received his B.S. degrees in physics and mathematics in 1999 and his M.S. degree in electrical engineering in 2001 from Vanderbilt University,

Nashville, Tennessee. He received his doctorate in applied mathematics and scientific computation in 2005 from the University of Maryland, College Park. He is currently an associate professor of biomedical engineering and electrical engineering at the University of Virginia, Charlottesville.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed. Lectures in Mathematics ETH Zürich. Basel: Birkhäuser Verlag, 2008.

[2] S. Angenent, S. Haker, and A. Tannenbaum, "Minimizing flows for the Monge–Kantorovich problem," *SIAM J. Math. Anal.*, vol. 35, no. 1, pp. 61–97, 2003.

[3] M. Arjovsky, S. Chintala, and L. Bottou. (2017). Wasserstein GAN. *arXiv*. [Online]. Available: https://arxiv.org/abs/1701.07875

[4] S. Basu, S. Kolouri, and G. K. Rohde, "Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 111, no. 9, pp. 3448–3453, 2014.

[5] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numer. Math.*, vol. 84, no. 3, pp. 375–393, 2000.

[6] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative Bregman projections for regularized transportation problems," *SIAM J. Sci. Comput.*, vol. 37, no. 2, pp. A1111–A1138, 2015.

[7] J.-D. Benamou, B. D. Froese, and A. M. Oberman, "Numerical solution of the optimal transportation problem using the Monge–Ampère equation," *J. Comput. Phys.*, vol. 260, pp. 107–126, 2014.

[8] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures," *J. Math. Imaging Vision*, vol. 51, no. 1, pp. 22–45, 2015.

[9] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Comm. Pure Appl. Math.*, vol. 44, no. 4, pp. 375–417, 1991.

[10] L. A. Caffarelli, "The regularity of mappings with a convex potential," *J. Am. Math. Soc.*, vol. 5, no. 1, pp. 99–104, Jan. 1992.

[11] R. Chartrand, K. Vixie, B. Wohlberg, and E. Bollt, "A gradient descent solution to the Monge-Kantorovich problem," *Appl. Math. Sci.*, vol. 3, no. 22, pp. 1071–1080, 2009.

[12] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," in *Machine Learning and Knowledge Discovery in Databases*. New York: Springer Berlin Heidelberg, 2014, pp. 274–289.

[13] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. (2016). Optimal transport for domain adaptation. *arXiv*. [Online]. Available: https://arxiv.org/abs/1507.00504

[14] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY: Curran Associates, 2013, pp. 2292–2300.

[15] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, J. M. Rodríguez-Rodríguez, "Tests of goodness of fit based on the L_2-Wasserstein distance," *Ann. Stat.*, vol. 27, no. 4, pp. 1230–1239, 1999.

[16] J. Delon, "Midway image equalization," *J. Math. Imaging Vision*, vol. 21, no. 2, pp. 119–134, 2004.

[17] T. A. El Moselhy and Y. M. Marzouk, "Bayesian inference with optimal maps," *J. Comput. Phys.*, vol. 231, no. 23, pp. 7815–7850, 2012.

[18] S. Ferradans, G.-S. Xia, G. Peyré, and J.-F. Aujol, "Static and dynamic texture mixing using optimal transport," in *Proc. Scale Space and Variational Methods in Computer Vision: 4th Int. Conf. (SSVM 2013)*, Leibnitz, Austria, June 2–6, 2013, pp. 137–148. doi: 10.1007/978-3-642-38267-3_12.

[19] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a Wasserstein loss," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY: Curran Associates, 2015, pp. 2044–2052.

[20] W. Gangbo and R. J. McCann, "The geometry of optimal transportation," *Acta Math.*, vol. 177, no. 2, pp. 113–161, 1996.

[21] E. Haber, T. Rehman, and A. Tannenbaum, "An efficient numerical method for the solution of the l_2 optimal mass transfer problem," *SIAM J. Sci. Comput.*, vol. 32, no. 1, pp. 197–211, 2010.

[22] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, "Optimal mass transport for registration and warping," *Int. J. Comput. Vision*, vol. 60, no. 3, pp. 225–240, 2004.

[23] L. V. Kantorovich, "On translation of mass" (in Russian), *Dokl. AN SSSR*, 37:199–201, 1942.

[24] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde. (2016). Transport-based analysis, modeling, and learning from signal and data distributions. *arXiv*. [Online]. Available: https://arxiv.org/abs/1609.04767

[25] S. Kolouri, S. R. Park, and G. K. Rohde, "The radon cumulative distribution transform and its application to image classification," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 920–934, 2016.

[26] S. Kolouri and G. K. Rohde. "Transport-based single frame super resolution of very low resolution face images," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 4876–4884.

[27] S. Kolouri, A. B. Tosun, J. A. Ozolek, and G. K. Rohde, "A continuous linear optimal transport approach for pattern analysis in image datasets," *Pattern Recognit.*, vol. 51, pp. 453–462, Mar. 2016.

[28] S. Kolouri, Y. Zou, and G. K. Rohde, "Sliced Wasserstein kernels for probability distributions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 5258–5267.

[29] R. Lai and H. Zhao. (2014). Multi-scale non-rigid point cloud registration using robust Sliced–Wasserstein distance via Laplace-Beltrami eigenmap. *arXiv*. [Online]. Available: https://arxiv.org/abs/1406.3758

[30] J. Lellmann, D. A. Lorenz, C. Schönlieb, and T. Valkonen, "Imaging with Kantorovich–Rubinstein discrepancy," *SIAM J. Imaging Sci.*, vol. 7, no. 4, pp. 2833–2859, 2014.

[31] B. Lévy, "A numerical algorithm for L_2 semi-discrete optimal transport in 3D," *ESAIM Math. Model. Numer. Anal.*, vol. 49, no. 6, pp. 1693–1715, 2015.

[32] P. Li, Q. Wang, and L. Zhang. "A novel earth mover's distance methodology for image matching with Gaussian mixture models," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 1689–1696.

[33] G. Loeper and F. Rapetti, "Numerical solution of the Monge–Ampère equation by a Newton's algorithm," *Comptes Rendus Math.*, vol. 340, no. 4, pp. 319–324, 2005.

[34] Q. Mérigot, "A multiscale approach to optimal transport," *Comput. Graph. Forum*, vol. 30, no. 5, pp. 1583–1592, 2011.

[35] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. Paris, France: De l'Imprimerie Royale, 1781.

[36] G. Montavon, K.-R. Müller, and M. Cuturi. (2015). Wasserstein training of Boltzmann machines. *arXiv*. [Online]. Available: https://arxiv.org/abs/1507.01972

[37] A. M. Oberman and Y. Ruan. (2015). An efficient linear programming method for Optimal Transportation. *arXiv*. [Online]. Available: https://arxiv.org/abs/1509.03668

[38] F. Otto, "The geometry of dissipative evolution equations: the porous medium equation," *Comm. Partial Differential Equations*, vol. 26, no. 1–2, pp. 101–174, 2001.

[39] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde, "Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning," *Med. Image Anal.*, vol. 18, no. 5, pp. 772–780, 2014.

[40] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde, "The cumulative distribution transform and linear pattern classification," *Appl. Comput. Harmonic Anal.*, 2017 in press.

[41] J. Rabin, S. Ferradans, and N. Papadakis. "Adaptive color transfer with relaxed optimal transport," in *Proc. 2014 IEEE Int. Conf. Image Processing (ICIP)*, 2014, pp. 4852–4856.

[42] A. Ramdas, N. Garcia, and M. Cuturi. (2015). On Wasserstein two sample testing and related families of nonparametric tests. *arXiv*. [Online]. Available: https://arxiv.org/abs/1509.02237

[43] G. K. Rohde, et al. Transport and other Lagrangian transforms for signal analysis and discrimination. [Online]. Available: http://faculty.virginia.edu/rohde/transport

[44] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[45] F. Santambrogio, *Optimal Transport for Applied Mathematicians*. New York: Birkhäuser, 2015.

[46] B. Schmitzer, "A sparse multiscale algorithm for dense optimal transport," *J. Math. Imaging Vision*, vol. 56, no. 2, pp. 238–259. doi: 10.1007/s10851-016-0653-9.

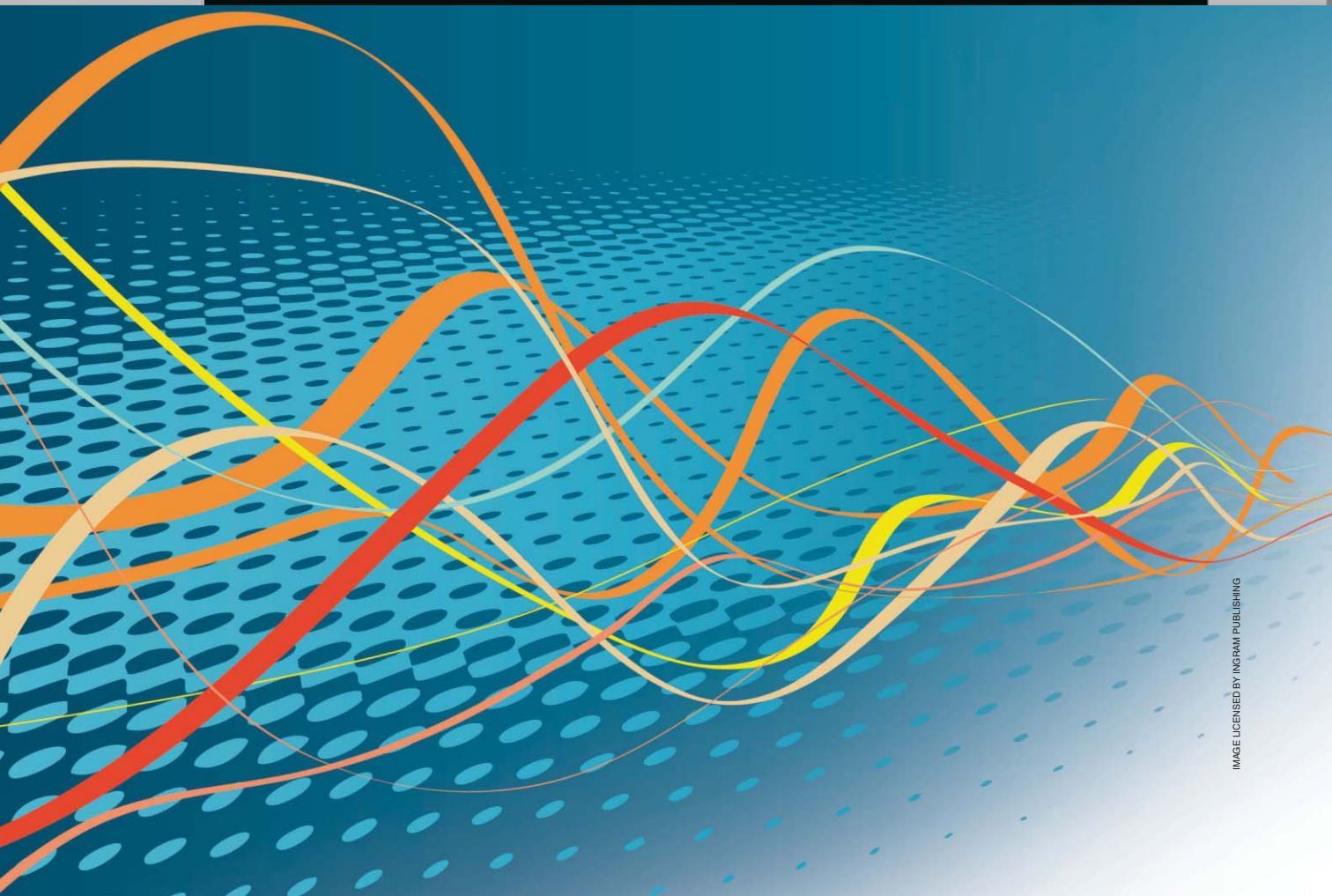
[47] J. Solomon, F. de Goes, P. A. Studios, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains," presented at the Special Interest Group on Computer Graphics and Interactive Techniques Conf., 2015.

[48] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. "Wasserstein propagation for semi-supervised learning," in *Proc. 31st Int. Conf. Machine Learning*, 2014, pp. 306–314.

[49] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin: Springer-Verlag, 2008.

[50] W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *Int. J. Comput. Vision*, vol. 101, no. 2, pp. 254–269, 2013.

SP



Adaptive Importance Sampling

The past, the present, and the future

Mónica F. Bugallo,
Víctor Elvira, Luca Martino,
David Luengo, Joaquín Míguez,
and Petar M. Djurić

Digital Object Identifier 10.1109/MSP.2017.2699226
Date of publication: 11 July 2017

A fundamental problem in signal processing is the estimation of unknown parameters or functions from noisy observations. Important examples include localization of objects in wireless sensor networks [1] and the Internet of Things [2]; multiple source reconstruction from electroencephalograms [3]; estimation of power spectral density for speech enhancement [4]; or inference in genomic signal processing [5]. Within the Bayesian signal processing framework, these problems are addressed by constructing posterior probability distributions of the unknowns. The posteriors combine optimally all of the information about the unknowns in the observations with the information that is present in their prior probability distributions. Given the posterior, one often wants to make inference about the unknowns, e.g., if we are estimating parameters, finding the values that maximize their posterior or the values that minimize some cost function given the uncertainty of the parameters.

Unfortunately, obtaining closed-form solutions to these types of problems is infeasible in most practical applications, and therefore, developing approximate inference techniques is of utmost interest.

A methodology that comes to the rescue for solving most difficult problems of inference is based on random drawing of samples. It was first applied systematically by the Italian physicist Enrico Fermi when he studied neutron diffusion [6]. However, no publication is available from him on this topic. Later, the methodology came to be known as *Monte Carlo* (*MC*) sampling.

The MC methods we know today were created by Stanislaw Ulam, John von Neumann, and others [7]. Their efforts coincided with the development of the first general computer and resulted in the Metropolis algorithm [8]. The next major advancement of MC methods came with a generalization of the Metropolis algorithm proposed by Hastings in 1970 [9]. All of these methods represent a family of simulation-based algorithms that aim at generating samples from a target probability distribution (often a posterior distribution in a Bayesian setting). The algorithms are based on constructing a Markov chain that has the desired distribution as its equilibrium distribution, which is why they are referred to as *Markov chain MC* (*MCMC*) *algorithms* [10] (a review of the history of MCMC sampling can be found in [7]). The most prominent MCMC algorithms remain the Metropolis–Hastings (MH) and Gibbs sampling algorithms [10]. Since the 1990s, MCMC-based methods have seen tremendous growth and success.

The MC methods we know today were created by Stanislaw Ulam, John von Neumann, and others.

Overview of importance sampling

An important alternative to MCMC sampling is the class of importance sampling (IS) methods. The IS methods are elegant, theoretically sound, simple to understand, and widely applicable [7]. Assume that the aim is to approximate a given target probability distribution. The basic IS mechanism consists of 1) drawing samples from simple proposal densities, 2) weighting the samples by accounting for the mismatch between the target and the proposal densities, and 3) performing the desired inference using the weighted samples. IS was first used in statistical physics for inference of rare events and, in particular, for estimating the probability of nuclear particles that penetrate shields [11]. Later, IS was also used as a variance reduction technique based on simulating from a proposal density instead of the target density [12]. The interest in IS techniques was running in parallel to the emergence of Bayesian computational methods. The interest was not only driven by their simplicity but also by their ability to estimate normalizing constants of the target distribution, a feature not shared by MCMC methods that turns out to be useful in many practical problems (e.g., model selection).

The performance of IS methods directly depends on the choice of the proposal densities [7]. When the method is applied naively, only few of the IS weights take relevant values, while the rest are negligible. This phenomenon is widely known in the IS literature as *weight degeneracy* [7]. If the goal is to estimate the mean of the samples of a target distribution, then the proposals must be adapted to parts of the space where the posterior probability is large, while if the focus is on a problem related to system reliability, then the probability of rare events is better approximated by placing the proposals in the tails of the posterior. Locating the regions from which samples should be drawn may not be easy, which suggests that the main challenge in implementing IS methods lies in finding good proposal densities. However, designing these proposals usually cannot be done a priori, and thus, adaptive procedures must be constructed and applied iteratively. The objective is that with passing iterations the quality of the samples improves, and the inference from them becomes more accurate. This leads us to the concept of adaptive IS (AIS). AIS methods are endowed with the nice feature of being able to learn from previously sampled values of the unknowns and, consequently, to become more accurate. It is important to note that the AIS algorithms must remain simple, i.e., both the drawing of samples and the computation of their weights should be easily managed.

In this article, we first go over the basics of IS and then proceed with explaining the learning process that takes place in AIS and with presenting several state-of-the-art methods. We discuss AIS estimators and their convergence properties and then show numerical results on signal processing examples. The article also provides an outlook of the research in AIS. For a clearer presentation, in Table 1 we display the notation used throughout the article.

Background (with examples)

Problem statement

Let us consider a generic inference problem in which a d_x-dimensional vector of unknown static real parameters, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, has a probability density function (pdf) given by

$$\tilde{\pi}(\mathbf{x}) = \frac{\pi(\mathbf{x})}{Z}, \tag{1}$$

where $\pi(\mathbf{x})$ is a nonnormalized nonnegative target function, and $Z = \int_{\mathcal{X}} \pi(\mathbf{x})\, d\mathbf{x}$ is a finite normalizing constant that may be unknown. The goal is to compute some particular moment of $\mathbf{x}$, which can be defined as

$$I = \int_{\mathcal{X}} f(\mathbf{x})\, \tilde{\pi}(\mathbf{x})\, d\mathbf{x}, \tag{2}$$

where $f(\cdot)$ can be any function of $\mathbf{x}$ that is integrable with respect to $\tilde{\pi}(\mathbf{x})$.

The previous mathematical formulation can be used to represent different problems, including the estimation of rare events [12] or Bayesian inference [7]. For instance, when estimating rare events, Z is perfectly known and the moment of

interest can be $f(\mathbf{x}) = \mathbb{I}_{g(\mathbf{x})>0}$, where $g(\mathbf{x})$ is a given function and $\mathbb{I}$ is the indicator function that takes the value 1 if $g(\mathbf{x}) > 0$ and 0 otherwise. In this case, $\tilde{\pi}(\mathbf{x})$ is completely characterized, and the challenge is in computing the integral given by (2). In Bayesian inference, $\tilde{\pi}(\mathbf{x})$ often represents the posterior distribution that is linked to some observed data, $\mathbf{y} \in \mathbb{R}^{d_y}$, and is expressed as

$$\tilde{\pi}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})}{Z(\mathbf{y})} \propto \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}), \tag{3}$$

Table 1. A summary of notation.

Notation	Description
d_x	Dimension of the unknown parameter vector
$\mathbf{x} \in \mathbb{R}^{d_x}$	Unknown realization of a parameter vector
d_y	Dimension of the observed data vector
$\mathbf{y} \in \mathbb{R}^{d_y}$	Observed data vector
i	Iteration variable
J	Total number of iterations
N	Number of proposal distributions in an iteration
K	Number of generated samples per proposal in an iteration
$\tilde{\pi}$	Target pdf
$\tilde{\pi}^K$	Approximated target pdf with K samples and weights
ℓ	Likelihood function
p_0	Prior distribution
Z	Normalizing constant
$\bar{I}^K$	Natural estimator computed from K samples generated from the target
$\hat{I}^K$	Nonnormalized estimator computed from K samples
$\tilde{I}^K$	Self-normalized estimator computed from K samples
$\mathbf{x}_{n,j}^{(k)}$	kth sample of the nth proposal at iteration j
$w_{n,j}^{(k)}$	IS weight associated with $\mathbf{x}_{n,j}^{(k)}$
$\bar{w}_{n,j}^{(k)}$	Normalized IS weight associated with $\mathbf{x}_{n,j}^{(k)}$
f	Test function/moment of the target
$q_{n,j}$	nth proposal function in the jth iteration
$\boldsymbol{\theta}_{n,j}$	Parameters defining the proposal $q_{n,j}$; e.g., $\theta_{n,j} = [\mu_{n,j}\ \mathbf{C}_{n,j}]$ for a Gaussian
$\boldsymbol{\mu}_{n,j}$	Location parameter (usually mean) of the proposal $q_{n,j}$
$\mathbf{C}_{n,j}$	Scale parameter (usually covariance) of the proposal $q_{n,j}$
$\rho_{n,j}$	Weight in the mixture of the nth proposal at iteration j
∇	Gradient
$\mathbf{H}_x$	Hessian evaluated at $\mathbf{x}$
λ_j	Gradient step at iteration j
$E_{\tilde{\pi}}[\cdot]$	Expected value with respect to the pdf $\tilde{\pi}$

where $p(\mathbf{x}|\mathbf{y})$ is the posterior pdf, $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $p_0(\mathbf{x})$ is the prior pdf, and $Z(\mathbf{y})$ is the model evidence or partition function. For some specific statistical models, e.g., when $p_0(\mathbf{x})$ is a conjugate prior of $\ell(\mathbf{y}|\mathbf{x})$ [13], $Z(\mathbf{y}) = \int_{\mathcal{X}} \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x})d\mathbf{x}$ can readily be obtained. In general, however, computing Z can be a very difficult problem. For this reason, we define the nonnormalized target function

$$\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})p_0(\mathbf{x}). \tag{4}$$

From here on and without loss of generality, we focus on the generic case, where $Z(\mathbf{y})$ is unknown. To simplify the notation, we drop the dependence of Z on $\mathbf{y}$ and write $Z \equiv Z(\mathbf{y})$. In the rest of the article, we refer to Z as a normalizing constant. This term is more general than model evidence or marginal likelihood, which are often used in Bayesian theory. Finally, we concentrate on real parameters and observations for the sake of clarity in the exposition. However, all of the AIS methods presented and the considerations performed throughout the article are directly applicable to multidimensional-complex target densities.

MC methods: motivation and basics

Obtaining closed-form solutions of the described problem is infeasible in most practical applications, and therefore, the next best thing is to develop approximate inference techniques with good accuracy. Let us assume that it is possible to draw K independent samples, $\{\mathbf{x}^{(k)}\}_{k=1}^K$, from the target distribution $\tilde{\pi}(\mathbf{x})$. The integral I can then be approximated by

$$\bar{I}^K = \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}^{(k)}), \qquad \text{where } \mathbf{x}^{(k)} \sim \tilde{\pi}(\mathbf{x}). \tag{5}$$

With the drawn samples, we can approximate the target probability distribution corresponding to the density $\tilde{\pi}(\mathbf{x})$ as

$$\tilde{\pi}^K(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}^{(k)}), \tag{6}$$

where $\delta(\mathbf{x} - \mathbf{x}^{(k)})$ is the Dirac delta function centered at $\mathbf{x}^{(k)}$. With this approximation, we can estimate I in (2) by

$$\begin{aligned} I &= \int_{\mathcal{X}} f(\mathbf{x})\tilde{\pi}(\mathbf{x})d\mathbf{x} \\ &\approx \int_{\mathcal{X}} f(\mathbf{x})\tilde{\pi}^K(\mathbf{x})d\mathbf{x} = \frac{1}{K}\sum_{k=1}^{K}\int_{\mathcal{X}} f(\mathbf{x})\delta(\mathbf{x} - \mathbf{x}^{(k)})d\mathbf{x}, \end{aligned} \tag{7}$$

which yields (5).

The estimator $\bar{I}^K$ is consistent with K, because it converges almost surely to I by the strong law of large numbers [7]. Moreover, it can be easily shown that the estimator is unbiased, i.e., $E_{\tilde{\pi}}[\bar{I}^K] = I$, and, assuming that $f(\mathbf{x})$ is real and square integrable, its variance is given by [7]

$$\mathrm{Var}_{\tilde{\pi}}(\bar{I}^K) = \frac{\mathrm{Var}_{\tilde{\pi}}(f(\mathbf{X}))}{K}. \tag{8}$$

This methodology is known as the *MC method* [7], and it was first described in [14].

As previously pointed out, very often, $\tilde{\pi}(\mathbf{x})$ does not have a known closed form, and it is not possible to draw samples from it. Moreover, in some other settings, it might not be convenient to generate samples from the target distribution even if it is possible. This is the case of rare-event estimation, where it is not efficient to simulate samples from $\tilde{\pi}(\mathbf{x})$ because the estimation of I would depend on a very low number of effective samples [15].

IS: motivation and basics

The IS methodology was first used in statistical physics for rare-event inference. More specifically, it was applied to estimate the probability of nuclear particles that penetrate shields [11]. Later, IS was also used as a variance reduction technique based on simulating from a proposal density instead of the target one, reducing the computational effort to compute rare events from the target distribution [12]. The interest in IS techniques has run in parallel to the growth of the theory of Bayesian inference. The reason for this is that often it is not possible to generate samples from the posterior distribution because it can only be evaluated up to a normalizing constant.

Let us consider K independent samples, $\{\mathbf{x}^{(k)}\}_{k=1}^{K}$, drawn from a single proposal pdf, $q(\mathbf{x})$, with heavier tails than the target, $\pi(\mathbf{x})$. Each sample has an associated importance weight given by

$$w^{(k)} = \frac{\pi(\mathbf{x}^{(k)})}{q(\mathbf{x}^{(k)})}, \quad k = 1, \ldots, K, \tag{9}$$

where the weights represent the significance of the samples in the approximation of the target by the considered proposal. Using the samples and weights, the integral in (2) can be approximated by a self-normalized estimator as

$$\tilde{I}^K = \frac{1}{K\hat{Z}} \sum_{k=1}^{K} w^{(k)} f(\mathbf{x}^{(k)}), \tag{10}$$

where $\hat{Z} = (1/K)\Sigma_{k=1}^{K} w^{(k)}$ is an unbiased estimator of $Z = \int_{\mathcal{X}} \pi(\mathbf{x}) d\mathbf{x}$ [7]. It is not difficult to see that now we approximate the target distribution by

$$\tilde{\pi}^K(\mathbf{x}) = \sum_{k=1}^{K} \bar{w}^{(k)} \delta(\mathbf{x} - \mathbf{x}^{(k)}), \tag{11}$$

where the $\bar{w}^{(k)}$s are normalized weights of the samples obtained by

$$\bar{w}^{(k)} = \frac{w^{(k)}}{\sum_{i=1}^{K} w^{(i)}}. \tag{12}$$

If the normalizing constant is known, then it is possible to use the nonnormalized estimator

$$\hat{I}^K = \frac{1}{KZ} \sum_{k=1}^{K} w^{(k)} f(\mathbf{x}^{(k)}). \tag{13}$$

Note that $\tilde{I}^K$ is only asymptotically unbiased, whereas $\hat{I}^K$ is unbiased. Both $\tilde{I}^K$ and $\hat{I}^K$ are consistent estimators of I, and their variance is directly related to the discrepancy between $\tilde{\pi}(\mathbf{x})|f(\mathbf{x})|$ and $q(\mathbf{x})$ [7]. However, when several different moments of the target must be estimated or the function f is unknown a priori, a common strategy in IS is to decrease the mismatch between the proposal $q(\mathbf{x})$ and the target $\tilde{\pi}(\mathbf{x})$ [16]. This is equivalent to minimizing the variance of the weights and, consequently, the variance of the estimator $\hat{Z}$.

Multiple IS: motivation and basics

The target density can only be evaluated pointwise, and therefore it cannot be easily characterized in many cases. This entails that finding a single good proposal pdf, $q(\mathbf{x})$, is not always possible. A robust alternative consists of using a set of proposal pdfs, $\{q_n(\mathbf{x})\}_{n=1}^{N}$. The resulting method is referred to as *multiple IS* (*MIS*), and it was greatly advanced during the 1990s in statistics and computer graphics simulation [12], [17], [18]. MIS constitutes the basis of most of the state-of-the-art AIS algorithms [19]–[24].

A general MIS framework has recently been proposed in which different sampling and weighting schemes can be combined [25]. Here, we briefly review the most common sampling and two common weighting schemes. Suppose that we draw one sample from each proposal pdf, i.e.,

$$\mathbf{x}_n \sim q_n(\mathbf{x}), \quad n = 1, \ldots, N, \tag{14}$$

where, because $K = 1$, we drop the superscript $^{(k)}$. The most common weighting strategies in the literature are

1) standard MIS (s-MIS) [19]:

$$w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}, \quad n = 1, \ldots, N. \tag{15}$$

2) deterministic mixture (DM) MIS (DM-MIS) [18]:

$$w_n = \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)} = \frac{\pi(\mathbf{x}_n)}{\frac{1}{N}\sum_{i=1}^{N} q_i(\mathbf{x}_n)}, \quad n = 1, \ldots, N, \tag{16}$$

where $\psi(\mathbf{x})$ represents the mixture pdf composed of all of the proposal pdfs evaluated at $\mathbf{x}$.

From the weighted set $\{\mathbf{x}_n, w_n\}_{n=1}^{N}$, generated by either the s-MIS or the DM-MIS methods described previously, we can compute a self-normalized estimator $\tilde{I}^N$ and a nonnormalized estimator $\hat{I}^N$ in the same way as in (10) and (13), respectively. The self-normalized $\tilde{I}^N$ is consistent and asymptotically unbiased, whereas the nonnormalized $\hat{I}^N$ is both consistent and unbiased. The DM approach is superior with respect to that of s-MIS in terms of variance of the estimator $\hat{I}^N$, as proved in [25]. Although both alternatives perform the same number of

target evaluations, the DM estimator is computationally more expensive with respect to the number of proposal evaluations. In particular, s-MIS and DM require N and N^2 evaluations, respectively. Therefore, in scenarios where the number of proposals N is large, the $O(N^2)$ in the number of proposal evaluations can be prohibitive. Alternative efficient solutions have recently been devised to mitigate this excess of computational load [26], [27].

All of the AIS methods presented and the considerations performed are directly applicable to multidimensional-complex target densities.

Figure 1 illustrates the processes of sampling and weighting based on the different methods explained in this section. More specifically, Figure 1(a) displays the generated samples and associated weights when sampling from the target distribution is possible. We observe that all of the weights are equal in this case. For both Figure 1(b) and (c), the generation of samples is performed using a single proposal pdf. However, the proposal pdfs, plotted with dashed lines, are differently located, and therefore one can appreciate how the second choice is more appropriate by observing the variability of the weight values. Note that the scale of the vertical axes is different to show the large weights in Figure 1(b). Figure 1(d) and 1(e) use the concept of MIS, i.e., there we use two proposal pdfs. The weights in Figure 1(d) are calculated using the standard formulation of weight update from (15), while in Figure 1(e), they are computed according to (16). It is clear that a smaller variance of the weights is achieved with the DM approach.

Finally, the validity of the possible different weighting schemes for MIS is justified in [25] by using the concept of a proper set of weighted samples. More precisely, the suitability of a particular MIS scheme is guaranteed if the nonnormalized estimator $\hat{I}^N$ and the normalizing constant estimator $\hat{Z}$ are unbiased and consistent, which also implies that the self-normalized estimator $\tilde{I}^N$ is consistent.

Adaptive importance sampling

The basics of AIS

The AIS methodology is based on an iterative process for gradual evolution of the single or multiple proposal densities

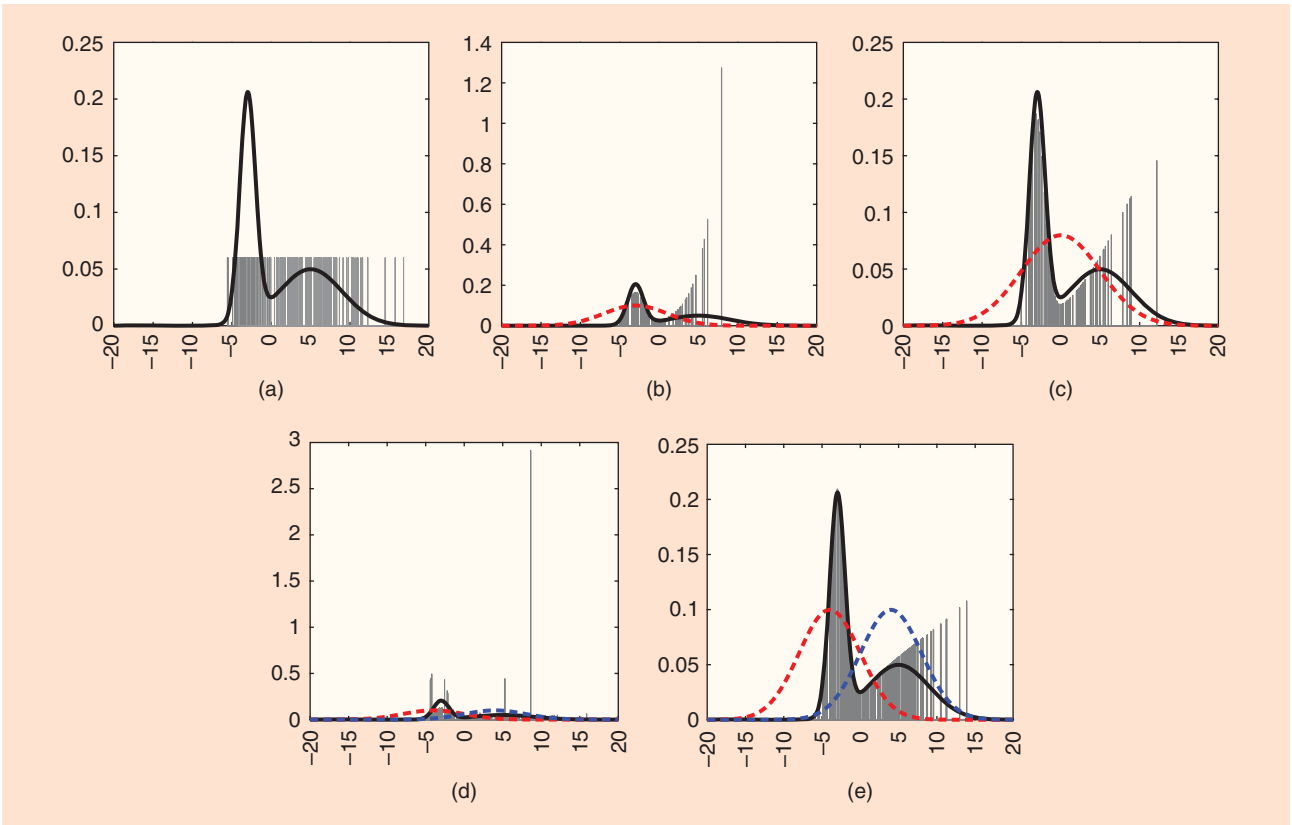


FIGURE 1. The approximations of the target pdf, $\pi(\mathbf{x})$, by different discrete probability distributions (displayed by thin bars with weights corresponding to heights of the bars). The target pdfs are shown by solid lines, while the proposal pdfs are plotted with dashed lines. (a) The MC sampling directly from the target, the ideal situation: an approximation with equally weighted samples, as they are drawn directly from the target. (b) IS, single-proposal pdf, and (c) IS, single-proposal pdf [with a better location than that in (b)]. (b) and (c) are approximations with IS and a single proposal to show the effect of the location: A better proposal placement leads to more uniform weights. (d) MIS with standard weights, and (e) MIS with DM weights. (d) and (e) are approximations with MIS and two proposals to show the effect of the choice of the weighting scheme: The DM approach leads to more uniform weights than the standard approach.

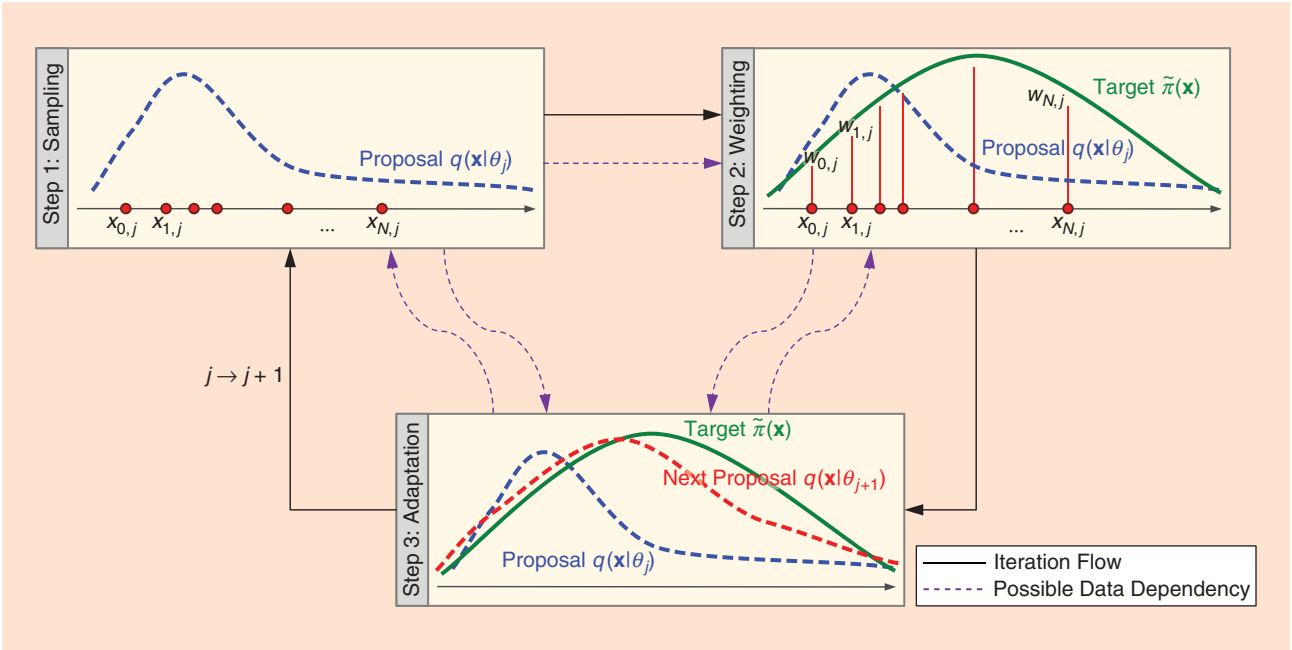


FIGURE 2. A generic flow diagram of the AIS methodology, showing the three steps that must be performed iteratively by any AIS algorithm (sampling, weighting, and adaptation) and the data flow among these steps.

to accurately approximate the target pdf. The procedure consists of three basic steps: generation of samples from a proposal or set of proposals (sampling), calculation of the importance of each of the samples (weighting), and updating (adapting) the parameters that define the proposal(s) to obtain the new proposal(s) for the next iteration. Figure 2 shows a simple flow diagram of the steps of AIS with only one proposal pdf. The diagram also shows the possible data dependencies among the basic steps.

In the general case, the algorithm is initialized with a set of N proposals $\{q_n(\mathbf{x}|\boldsymbol{\theta}_{n,1})\}_{n=1}^{N}$, each one parameterized by a vector $\boldsymbol{\theta}_{n,1}$. After drawing a set of samples, $\mathbf{x}_{n,1}^{(k)}$, $n = 1,\ldots,N, k = 1,\ldots,K$ (recall that K is the number of samples generated by a proposal), and weighting them, one obtains a discrete probability distribution that approximates the target distribution, $\{x_{n,1}^{(k)}, w_{n,1}^{(k)}\}$, $n = 1,\ldots,N, k = 1,\ldots,K$. Then, the parameters of the nth proposal are updated from $\boldsymbol{\theta}_{n,1}$ to $\boldsymbol{\theta}_{n,2}$. This process is repeated, i.e., sampling, weighting, and moving from $\boldsymbol{\theta}_{n,j}$ to $\boldsymbol{\theta}_{n,j+1}$, until an iteration stoppage criterion is met (e.g., a maximum number of iterations, J, is reached). Table 2 outlines the main steps of the general algorithm.

Figure 3 shows the evolution in the approximation of a target pdf, $\tilde{\pi}(\mathbf{x})$, which in this case is a mixture of two Gaussian pdfs. In this example just one Gaussian proposal ($N = 1$) is used, $q_1(\mathbf{x})$, with initial vector parameter $\boldsymbol{\theta}_{1,1} = [\mu_1\ \sigma_1^2] = [-4\ \ 3]$, where μ_1 and σ_1^2 denote the mean and the variance, respectively. Figure 3 displays three iterations of the AIS algorithm, where the initial parameter vector $\boldsymbol{\theta}_{1,1}$ is updated in the next proposal so that it can produce samples and weights that yield a better approximation of the target distribution. Note that the final scale and location of the proposal is much more adequate than the starting proposal in that it effectively covers both modes of the target.

Table 2. The generic AIS algorithm.

Initialization

Choose K, N, J, $\{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N}$

For $j = 1,\ldots,J$:

1) Sampling

Draw K samples from each of the N proposal pdfs, $\{q_{n,j}(\boldsymbol{\theta}_{n,j})\}_{n=1}^{N}$, $\mathbf{x}_{n,j}^{(k)}$, $k = 1,\ldots,K, n = 1,\ldots,N$.

2) Weighting

Calculate the weights, $w_{n,j}^{(k)}$, for each of the generated KN samples.

3) Adaptation

Update the proposal parameters $\{\boldsymbol{\theta}_{n,j}\}_{n=1}^{N} \rightarrow \{\boldsymbol{\theta}_{n,j+1}\}_{n=1}^{N}$.

Outputs

Return the KNJ pairs $\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$ for all $k = 1,\ldots,K, n = 1,\ldots,N, j = 1,\ldots,J$.

To approximate the integral I in (2), there exist different possibilities for combining all of the KNJ weighted samples, $\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$, generated by the AIS method [28]. A common (and straightforward) choice is to assign to each sample a normalized weight $\bar{w}_{n,j}^{(k)}$, which considers all of the weights, i.e.,

$$\bar{w}_{n,j}^{(k)} = \frac{w_{n,j}^{(k)}}{\sum_{l=1}^{J}\sum_{i=1}^{N}\sum_{r=1}^{K} w_{i,l}^{(r)}}. \tag{17}$$

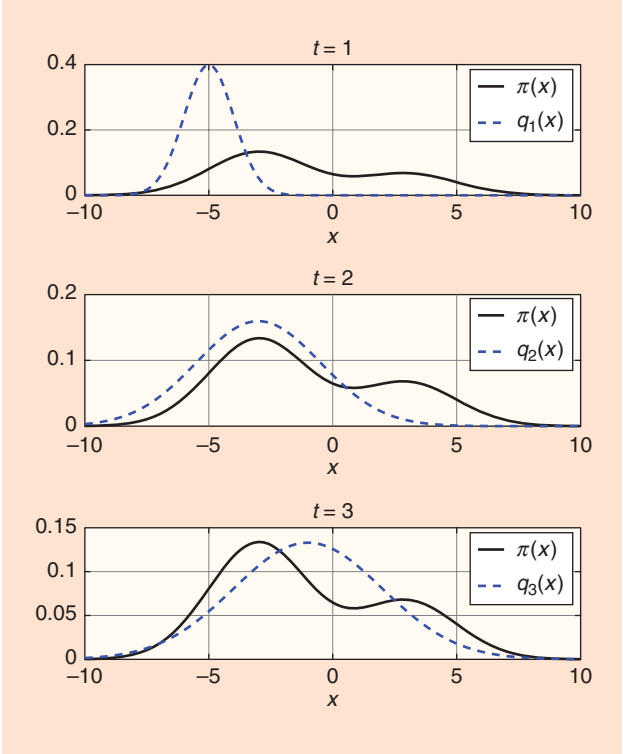


FIGURE 3. A proposal adaptation through AIS. The initial proposal $q_1(x)$ (too narrow and poorly placed) is iteratively moved toward a better location at some intermediate location between the two modes of the target pdf and widened to properly cover the effective support of the target.

Hence, the self-normalized AIS estimator is $\tilde{I}^{KNJ} = \sum_{j=1}^{J} \sum_{n=1}^{N} \sum_{k=1}^{K} \bar{w}_{n,j}^{(k)} f(\mathbf{x}_{n,j}^{(k)})$.

Modern AIS methods

AIS methods got their turn in the spotlight of MC computations after the publication of the population MC (PMC) sampling method by Cappé et al. in 2004 [19], notwithstanding the existence of several AIS schemes at that time (see [28] for a review). The PMC methodology offered a framework to adapt a population of proposals that was simple, flexible, and free from the convergence and ergodicity issues of adaptive MCMC techniques. The original PMC algorithm used a multinomial resampling stage (note that any of the better alternative resampling strategies developed for particle filters can also be used [29]) and was unstable due to the use of the s-MIS weighting strategy of (15). However, the proposed approach raised a considerable interest within the computational statistics community, and improved PMC algorithms shortly followed, like the D-kernel PMC [30], [31] or the mixture PMC (M-PMC) [20]. Furthermore, several authors have recently shown that the performance of PMC can be improved even more through the use of a nonlinear transformation of the weights [32] or the combination of the DM weighting scheme of (16) and sophisticated resampling schemes [24].

On the other hand, encouraged by the renewed interest in AIS methods spurred by the PMC approach, several authors have proposed AIS algorithms that do not fall within the PMC framework. For instance, the idea of incremental IS mixtures [originally proposed in (33)] was taken up again by Cornuet et al. in the adaptive MIS (AMIS) method [21]. AMIS uses a single proposal per iteration, but applies the DM weighting scheme of (16) using a mixture composed of the present and all past proposal pdfs. Much more robust and stable estimators are thus obtained, but at the expense of a substantial increase in the computational cost. An alternative to AMIS is the recently proposed adaptive population IS (APIS) algorithm [22]. APIS is also based on the DM weighting scheme of (16), but it uses a mixture with a fixed number of proposals per iteration. In this way, APIS inherits the robustness and stability of AMIS but with the benefit of allowing a user-controllable computational cost that does not increase as the algorithm is iterated. Moreover, gradient information can be incorporated to the APIS algorithm to improve the performance in high-dimensional state spaces [34].

Finally, note that the combination of MCMC and AIS techniques has also been considered in several works. For instance, MCMC steps can be used to accelerate the adaptation of the AIS technique [22], or the MCMC outputs can be used to build a proposal distribution for AIS estimation [35]. Sequential MC samplers have also been suggested as AIS schemes in static scenarios [36].

Implementation and classification of AIS algorithms

Implementation of AIS algorithms

Many important AIS algorithms have been proposed in the literature in the last two decades. In this section, we describe in detail some of the most popular AIS algorithms.

- *Standard PMC* [19]: In this algorithm, N proposals are adapted via resampling, which is a well-known mechanism in MC methodologies that allows us to select the most promising samples and to eliminate those with low weights to avoid particle degeneracy [29]. At each iteration, exactly one sample is drawn from each proposal and weighted with the standard IS weights calculated by (15). Then, N multinomial resampling steps (with replacement) are performed within the population of the N drawn samples (one sample is generated per proposal, i.e., $K = 1$). The surviving set of particles constitutes the set of location parameters for the next population of proposals.
- *M-PMC* [20]: For this method, the proposal used to generate K samples at each iteration is a mixture of N kernels, where the mixture is adapted to decrease the Kullback–Leibler (KL) divergence between the mixture and the target. In its simplest version, the algorithm adapts the location, scale, and weight of each kernel in the mixture.
- *Nonlinear PMC* (*N-PMC*) [32]: In this algorithm, the weights are computed in two steps. First, standard importance weights $w_j^{(k)}$ are obtained. Then, a nonlinear function is applied to calculate a set of transformed weights $\check{w}_j^{(k)}$. The goal of this transformation is to reduce the variance of the weights and avoid, or at least mitigate, the

weight degeneracy problem. While the standard weights can be used for estimation, the nonlinearly transformed weights are crucially used for the adaptation step. The latter can be carried out in different ways, with [32] advocating for a simple Gaussian proposal where both the mean vector and the covariance matrix are adapted through the iterations.

- *Layered AIS* (*LAIS*) [23]: The adaptive process of the LAIS algorithm is independent of the samples drawn at each iteration. In particular, the algorithm can be seen as a two-layer procedure in which the location parameters of the proposals are adapted through one or several MCMC steps with the target as the stationary distribution. In its basic version, a single MCMC step is independently performed at each location parameter.
- *DM-PMC* [24]: This algorithm meets the simplicity of the standard PMC of [19] with a very high performance. DM-PMC calculates the weights using (16) instead of (15), which provides two important advantages, specifically, the variance of the estimators is decreased (see [25]) and the resampling step with the DM weights promotes the replication of proposals in relevant parts of the target that are underrepresented by the set of proposals (i.e., the exploration is coordinated). DM-PMC generates K samples per each of the N proposals (instead of one, as in [19]). At each iteration, the population of KN samples must be reduced to N via either global or local resampling (LR).
- *AMIS* [21]: In this algorithm, just one proposal is used and adapted over the iterations. The adaptive procedure consists of estimating the moments of the target with the available set of K weighted samples and fitting the moments of the proposal. Its key feature is the reweighting of all of the past samples with a temporal mixture weight where the whole sequence of proposals is used in the denominator.
- *Gradient APIS* (*GAPIS*) [34]: Similar to the LAIS algorithm, GAPIS adapts N proposals by a process that is independent of the samples. In its basic version, the location parameters of the proposals are adapted via a gradient ascent of the target and the scale parameter by using the Hessian of the target. An advanced implementation is proposed that adds a repulsive interaction among proposals to promote a cooperative exploration of the target.

In Tables 3 and 4, six out of the seven previous algorithms are outlined by means of pseudocodes. Note that we follow

Table 3. The pseudocodes of PMC, DM-PMC, and LAIS.

PMC	DM-PMC	LAIS
	Initialization	
$J, N, K = 1$ $\{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N} \equiv \{\boldsymbol{\mu}_{n,1}, \mathbf{C}_n\}_{n=1}^{N}$	$J, N, K,$ $\{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N} \equiv \{\boldsymbol{\mu}_{n,1}, \mathbf{C}_n\}_{n=1}^{N}$	$J, N, K,$ $\{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N} \equiv \{\boldsymbol{\mu}_{n,1}, \mathbf{C}_n\}_{n=1}^{N}$
For $j = 1, \ldots, J$:		
	1) Sampling	
$\mathbf{x}_{n,j} \sim q_{n,j}(\mathbf{x} \mid \boldsymbol{\mu}_{n,j}, \mathbf{C}_n)$ $n = 1, \ldots, N$	$\mathbf{x}_{n,j}^{(k)} \sim q_{n,j}(\mathbf{x} \mid \boldsymbol{\mu}_{n,j}, \mathbf{C}_n)$ $n = 1, \ldots, N$ $k = 1, \ldots, K$	$\mathbf{x}_{n,j}^{(k)} \sim q_{n,j}(\mathbf{x} \mid \boldsymbol{\mu}_{n,j}, \mathbf{C}_n)$ $n = 1, \ldots, N$ $k = 1, \ldots, K$
	2) Weighting	
$w_{n,j} = \dfrac{\pi(\mathbf{x}_{n,j})}{q_{n,j}(\mathbf{x}_{n,j})}$ $n = 1, \ldots, N$	$w_{n,j}^{(k)} = \dfrac{\pi(\mathbf{x}_{n,j}^{(k)})}{\frac{1}{N}\sum_{i=1}^{N} q_{i,j}(\mathbf{x}_{n,j}^{(k)})}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$	$w_{n,j}^{(k)} = \dfrac{\pi(\mathbf{x}_{n,j}^{(k)})}{\frac{1}{N}\sum_{i=1}^{N} q_{i,j}(\mathbf{x}_{n,j}^{(k)})}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$
	3) Adaptation	
Multinomial resampling with replacement over $\{\mathbf{x}_{n,j}, \bar{w}_{n,j} = \dfrac{w_{n,j}}{\sum_{i=1}^{N} w_{i,j}}\}_{n=1}^{N}$ to update $\{\mu_{n,j+1}\}_{n=1}^{N}$.	Multinomial resampling with replacement over $\{\mathbf{x}_{n,j}^{(k)}, \bar{w}_{n,j}^{(k)} = \dfrac{w_{n,j}^{(k)}}{\sum_{j=1}^{N}\sum_{m=1}^{K} w_{i,j}^{(m)}}\}_{n=1,k=1}^{N,K}$ to update $\{\mu_{n,j+1}\}_{n=1}^{N}$.	One (or more) MCMC steps from $\mu_{n,j}$ to $\mu_{n,j+1}$, with $\tilde{\pi}$ as a stationary distribution, for $n = 1, \ldots, N$.
	Outputs	
$\{\mathbf{x}_{n,j}, w_{n,j}\}$ $n = 1, \ldots, N$ $j = 1, \ldots, J$	$\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$ $j = 1, \ldots, J$	$\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$ $j = 1, \ldots, J$

Table 4. The pseudocodes of AMIS, GAPIS, and M-PMC.

AMIS	GAPIS	M-PMC
	Initialization	
$J, K, N = 1,\ \boldsymbol{\theta}_1 \equiv \{\boldsymbol{\mu}_1, \mathbf{C}_1\}$ **For** $j = 1, \ldots, J$:	$J, N, K,\ \{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N} \equiv \{\boldsymbol{\mu}_{n,1}, \mathbf{C}_n\}_{n=1}^{N}$	$J, N, K,\ \{\boldsymbol{\theta}_{n,1}\}_{n=1}^{N} \equiv \{\rho_{n,1}, \boldsymbol{\mu}_{n,1}, \mathbf{C}_{1,n}\}_{n=1}^{N}$
	1) Sampling	
$\mathbf{x}_j^{(k)} \sim q_j(\mathbf{x}\vert\boldsymbol{\mu}_j, \mathbf{C}_j)$ $k = 1, \ldots, K$	$\mathbf{x}_{n,j}^{(k)} \sim q_{n,j}(\mathbf{x}\vert\boldsymbol{\mu}_{n,j}, \mathbf{C}_n)$ $n = 1, \ldots, N$ $k = 1, \ldots, K.$	$\mathbf{x}_j^{(k)} \sim \sum_{i=1}^{N} \rho_{i,j} q_{i,j}(\mathbf{x}\vert\boldsymbol{\mu}_{i,j}, \mathbf{C}_{i,j})$, $k = 1, \ldots, K.$
	2) Weighting	
$w_j^{(k)} = \dfrac{\pi(\mathbf{x}_j^{(k)})}{\frac{1}{j}\sum_{i=1}^{j} q_i(\mathbf{x}_j^{(k)})}$ $k = 1, \ldots, K$	$w_{n,j}^{(k)} = \dfrac{\pi(\mathbf{x}_{n,j}^{(k)})}{\frac{1}{N}\sum_{i=1}^{N} q_{i,j}(\mathbf{x}_{n,j}^{(k)})}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$	$w_j^{(k)} = \dfrac{\pi(\mathbf{x}_j^{(k)})}{\sum_{i=1}^{N} \rho_{i,j} q_{i,j}(\mathbf{x}_j^{(k)})}$ $k = 1, \ldots, K$
	3) Adaptation	
Update $\boldsymbol{\mu}_{j+1}$ and $\mathbf{C}_{j+1}$ with the empirical mean and covariance using all of the weighted samples.	Use a suitable λ_j to update $\boldsymbol{\mu}_{n,j+1} = \boldsymbol{\mu}_{n,j} + \lambda_j \nabla \log(\pi(\boldsymbol{\mu}_{n,j}))$ and the Hessian matrix of $-\log(\pi(\mathbf{x}))$ to update $\mathbf{C}_{n,j+1} = (\mathbf{H}_{\mu_{n,j}})^{-1}$.	Update $\{\rho_{n,j+1}, \boldsymbol{\mu}_{n,j+1}, \mathbf{C}_{n,j+1}\}_{n=1}^{N}$ by minimizing the KL distance between the proposal and the target approximation.
	Outputs	
$\{\mathbf{x}_j^{(k)}, w_j^{(k)}\}$ $k = 1, \ldots, K$ $j = 1, \ldots, J$	$\{\mathbf{x}_{n,j}^{(k)}, w_{n,j}^{(k)}\}$ $n = 1, \ldots, N$ $k = 1, \ldots, K$ $j = 1, \ldots, J$	$\{\mathbf{x}_j^{(k)}, w_j^{(k)}\}$ $k = 1, \ldots, K$ $j = 1, \ldots, J$

the structure sampling, weighting, and adaptation described in Figure 2 and Table 2. We have skipped the N-PMC scheme in these tables for the sake of clarity. We simply point out that, in this algorithm, the standard weights $w_{n,j}^{(k)}$ are transformed using a nonlinearity Φ, e.g., $\breve{w}_{n,j}^{(k)} = \Phi\left(k, \{w_{n,j}^{(l)}\}_{l=1}^{K}\right)$. These transformed weights are then fed to the adaptation stage. In [32], the nonlinearity $\Phi(\cdot,\cdot)$ is either a tempering or a simple truncation of the largest weights, while the adaptation is carried out as in the AMIS method of Table 4.

Classification of relevant AIS algorithms

Table 5 serves as a summary and compares the main features of different AIS implementations. The features include the number of proposals, the weighting procedure, the updating strategy of the parameters, and the updated parameters. Note that most of the algorithms use more than one proposal. However, due to the adaptive procedure, even with $N = 1$, more than one proposal is used. This is exploited in AMIS and in some implementations of LAIS, where the temporal mixture of proposals is used to reweight the samples via DM IS weights. Note that the different adaptive mechanisms can be classified into a mechanism based on 1) resampling, 2) moment matching, and 3) independent adaptive processes. Moreover, the moment matching can include all of the past weighted samples (AMIS) or just those of the current iteration (APIS). Figure 4 shows three possible dependence charts related to generated samples and the adaptation of the proposal parameters. Note also that, although all of the proposal parameters can be adapted, in the basic implementation of most algorithms, just the location parameters are adapted.

The interest in IS techniques was not only driven by their simplicity but also by their ability to estimate normalizing constants of the target distribution.

Table 6 provides a comparison of the computational complexity of the different algorithms. We display the number of target and proposal evaluations and also the same quantities per drawn sample. We observe that in AMIS, the number of proposal evaluations is increased with the number of iterations, while in the algorithms with DM weights, this problem appears when we increase the number of proposals. In the latter case, the strategies proposed in [26] and [27] can be employed to reduce the number of proposal evaluations. Although this is not displayed in Table 6, the GAPIS algorithm also requires

Table 5. A comparison of various AIS algorithms according to different features.

Algorithm	# Proposals	Weighting	Adaptation Strategy	Parameters Adapted
Standard PMC	$N > 1$	Standard	Resampling	Location
M-PMC	$N > 1$	Spatial mixture	Resampling	Location
N-PMC	Either	Nonlinear	Moment estimation	Location/scale
LAIS	$N > 1$	Generic mixture	MCMC	Location
DM-PMC	$N > 1$	Spatial mixture	Resampling	Location
AMIS	$N = 1$	Temporal mixture	Moment estimation	Location/scale
GAPIS	$N > 1$	Spatial mixture	Gradient process	Location/scale
APIS	$N > 1$	Spatial mixture	Moment estimation	Location

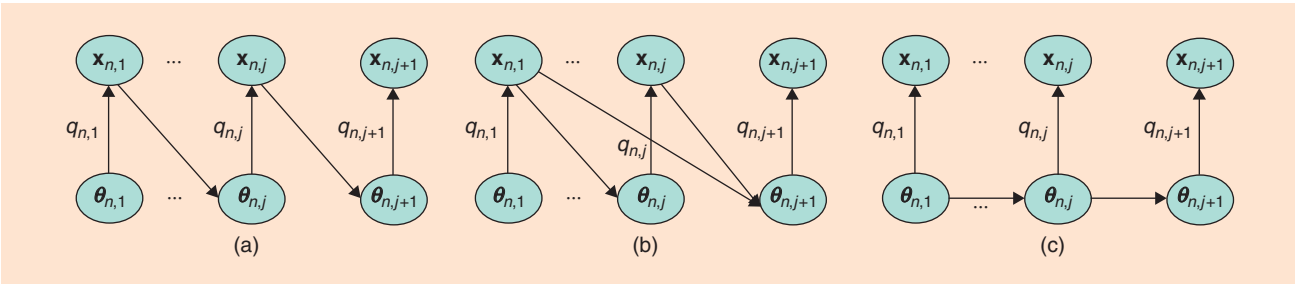


FIGURE 4. A graphical description of three possible dependencies between the adaptation of the proposal parameters $\boldsymbol{\theta}_{n,t}$ and the samples. Note that $q_{n,t} \equiv q_{n,t}(\mathbf{x}|\boldsymbol{\theta}_{n,t})$. (a) The proposal parameters are adapted using the last set of drawn samples (standard PMC, DM-PMC, N-PMC, M-PMC, APIS). (b) The proposal parameters are adapted using all drawn samples up to the latest iteration (AMIS). (c) The proposal parameters are adapted using an independent process from the samples (LAIS, GAPIS).

NJ radient and Hessian evaluations in total, i.e., one per proposal at each iteration.

A brief summary and comparison of AIS algorithms

In this section, we provide intuition behind the relevant AIS algorithms presented previously. The standard PMC [19] opened the door for the fast growth of the AIS methodology. While the simplicity is its main advantage, the use of the standard IS weights of (15) has two adverse effects: 1) the variance of the estimators is increased, and 2) each importance weight measures the difference between the target and a specific proposal (regardless of where the other $N-1$ proposals are placed). The latter effect precludes a stable and coordinated adaptation of the whole mixture of proposals and provokes a path degeneracy due to the resampling step.

The M-PMC [20] addresses the weak points of the standard PMC by applying a robust Rao-Blackwellization step in the adaptation of the proposals. The goal in M-PMC is to iteratively decrease the KL divergence between the target and the mixture of proposals (for the first time, they are seen as a mixture instead of a collection of proposals). M-PMC is more robust and allows for the adaptation of the covariance of each proposal and its weight in the mixture. The disadvantage is the extra computational cost and the potential instability in the adaptation of the covariance (it can tend to a delta) and in the mixture weights (the mixture can end up being formed by just one proposal).

Table 6. A comparison of various AIS algorithms according to the computational complexity.

Algorithm	Number of Target Evaluations	Number of Proposal Evaluations	Number of Target Evaluations/ Sample	Number of Proposal Evaluations/ Sample
Standard PMC	NJ	NJ	1	1
N-PMC	NJ	NJ	1	1
M-PMC	KJ	KNJ	1	N
LAIS	$K(N+1)J$	KN^2J	$1+1/N$	N
DM-PMC	KNJ	KN^2J	1	N
AMIS	KJ	KJ^2	1	J
GAPIS	KNJ	KN^2J	1	N
APIS	KNJ	KN^2J	1	N

The DM-PMC addresses the open challenges of the standard PMC in a different way. The use of DM IS weights, followed by the resampling step, implicitly aims at iteratively reducing the mismatch between the target and the mixture of proposals [see (16)]. In addition, DM-PMC allows to draw $K > 1$ samples per proposal per iteration, which improves the local exploration in the region of each proposal and then

increases the stability of the algorithm. Two variants of the algorithm, global resampling (GR)-PMC and LR-PMC, allow for different resampling steps to transition from NK samples in iteration j to N proposals in iteration $j+1$. The advantage of DM-PMC and its variants is the simplicity in the implementation and the high performance. The disadvantage is that only the location parameters of the proposals are adapted.

In general, all of the PMC-based algorithms use the set of weighted samples to adapt the proposals. While this recycling is efficient, the interdependence between the samples and the next generation of proposals hinders the theoretical analysis of the algorithms.

The LAIS algorithm disconnects the sampling and the adaptive procedures by establishing a two-layer scheme [see Figure 4(c)]. In its simplest version, the adaptive layer of LAIS is driven by MH chains, enjoying some of the advantages of the MCMC methods, e.g., their good behavior in high dimension. The LAIS scheme is simple and shows good performance, but again, it does not adapt the covariance of the proposals.

The GAPIS algorithm also decouples the adaptation and sampling procedures, adding the information of the gradient and Hessian of the target in the adaptation of the proposals. This scheme performs well in challenging problems, even in high dimensions, and is able to adapt the location and scale parameters of the proposals. Its main disadvantage is the complexity associated with the computation of the gradient and the Hessian.

The AMIS algorithm is also simple because the proposal adaptation is carried out via moment matching. The algorithm has shown good performance in a variety of problems. Furthermore, it is robust because the IS weights are permanently recomputed via Rao-Blackwellization by using the DM idea with the mixture of temporal proposals. The main disadvantage is precisely this recomputation of all of the weights at every iteration, which precludes its use when the needed number of iterations J is high. The DM-PMC, LAIS, and GAPIS methods are particularly well suited to multimodal target distributions, which are often hard for conventional algorithms (e.g., nonadaptive importance samplers or classical MCMC schemes).

Finally, note that the nonlinear transformation of the importance weights featured by the N-PMC method (to reduce the weight variance) can readily be applied to other schemes (DM-PMC, AMIS, etc.). This is especially useful at the first stages of the adaptation, when the proposal(s) can still be poorly aligned with the target density, and the use of transformed weights can often prevent severe sample impoverishment. Once the proposal is roughly adapted, the nonlinear transformation can be dropped and conventional weights can be used to reduce the computational cost.

Discussion of AIS methods

Convergence of IS estimators

The convergence of IS schemes is often assessed in terms of the approximation of integrals of test functions. Specifically, if $\mathbf{X}$ is a random vector of interest, taking values on $\mathbb{R}^{d_x}$ and with pdf $\tilde{\pi}(\mathbf{x})$, then we study the approximation of the integral

$$I(f) = \int_{\mathcal{X}} f(\mathbf{x})\tilde{\pi}(\mathbf{x})d\mathbf{x}, \tag{18}$$

where $f: \mathbb{R}^{d_x} \to \mathbb{R}$ is a real test function, assumed integrable with respect to the density $\tilde{\pi}(\mathbf{x})$ (now we make the test function f explicit in the notation of the integral). Note that $I(f)$ is the expected value of the real random variable $f(\mathbf{X})$, which can be alternatively denoted by $E_{\tilde{\pi}}[f(\mathbf{X})]$, and the integrability assumption simply states that this expectation exists, i.e., $E_{\tilde{\pi}}[f(\mathbf{X})] < \infty$.

A standard IS scheme with a proposal function $q(\mathbf{x})$ produces a set of random weighted samples $\{\mathbf{x}^{(k)}, w^{(k)}\}_{k=1}^{N}$, where $\mathbf{x}^{(k)} \sim q(\mathbf{x})$ and $w^{(k)} = \pi(\mathbf{x}^{(k)})/q(\mathbf{x}^{(k)})$, that we use to approximate the integral $I(f)$ as

$$\tilde{I}^K(f) = \frac{1}{\sum_{i=1}^{K} w^{(i)}} \sum_{k=1}^{K} w^{(k)} f(\mathbf{x}^{(k)}). \tag{19}$$

Note that $\tilde{I}^K(f)$ is a random variable itself. Intuitively, we expect that the error $I(f) - \tilde{I}(f)$ should vanish, in some proper probabilistic sense, when $K \to \infty$. This is, indeed, a consequence of the strong law of large numbers [7]. Assuming that $q(\mathbf{x}) > 0$ whenever $\pi(\mathbf{x}) > 0$, it can be proved that [37]

$$\lim_{K\to\infty} \tilde{I}^K(f) = I(f) \quad \text{almost surely (a.s.)}, \tag{20}$$

which implies that $\tilde{I}^K(f)$ is a consistent estimator of $I(f)$. Under additional, yet mild, assumptions on the weight and test functions, e.g.,

$$E_{\tilde{\pi}}[w(\mathbf{X})] < \infty \quad \text{and} \quad E_{\tilde{\pi}}[f^2(\mathbf{X})w(\mathbf{X})] < \infty, \tag{21}$$

a central limit theorem (CLT) also holds for the IS estimator [37]. [Note that here we use the notation $w(\mathbf{X})$ to remind the reader that the weights are functions of the random vector $\mathbf{X}$ and therefore are random variables themselves.] In particular,

$$\sqrt{K}(\tilde{I}^K(f) - I(f)) \overset{d}{=} \mathcal{N}(0, \sigma^2(f)), \tag{22}$$

where $\overset{d}{=}$ denotes convergence of the limit in distribution and the limit variance depends on the test function, namely, $\sigma^2(f) \propto E_{\tilde{\pi}}[(f(\mathbf{X}) - E_{\tilde{\pi}}[f(\mathbf{X})])^2 w(\mathbf{X})]$.

Equation (22) is one of various results that show how IS estimators converge with the optimal MC rate $O(1/\sqrt{K})$, i.e., the errors are asymptotically of the same order as with the standard MC estimator constructed with K independent identically distributed samples from the target pdf $\tilde{\pi}(\mathbf{x})$. The same optimal rate is obtained for the convergence of the L_p norms of the errors $\tilde{I}^K(f) - I(f)$ if we assume that both the test function f and the weight function w are bounded, specifically,

$$\|f\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^{d_x}} |f(\mathbf{x})| < \infty \quad \text{and}$$
$$\|w\|_\infty = \sup_{\mathbf{x}\in\mathbb{R}^{d_x}} |w(\mathbf{x})| = \sup_{\mathbf{x}\in\mathbb{R}^{d_x}} \left|\frac{\pi(\mathbf{x})}{q(\mathbf{x})}\right| < \infty, \tag{23}$$

where $\|Z\|_p$ indicates the L_p norm of the random variable Z with a pdf $g(z)$, i.e., $\|Z\|_p = \left(\int Z^p g(z)dz\right)^{(1/p)}$. Whenever (23) holds, it can be proved that [38]

$$\|I(f) - \tilde{I}^K(f)\|_p \le \frac{c\|f\|_\infty}{\sqrt{K}}, \tag{24}$$

for any $p \ge 1$ and some constant $c < \infty$ independent of K. The inequality in (24) is easily extended, using a standard argument based on the Markov inequality and the Borel–Cantelli lemma [39], to yield $\lim_{K\to\infty} \tilde{I}^K(f) = I(f)$ a.s.

A more sophisticated analysis allows us to obtain an upper bound for the random error (not just for its L_p norm) of the form [38]

$$|I(f) - \tilde{I}^K(f)| \le \frac{U_\epsilon}{K^{\frac{1}{2}-\epsilon}}, \tag{25}$$

where $\epsilon \in (0,(1/2))$ is an arbitrarily small constant and U_ϵ is an almost surely finite random variable independent of K. The inequality (25) holds for every value of K, hence it is stronger than the classical CLT of (22). As (22), it displays the optimal MC error rate $O(1/\sqrt{K})$, because $\epsilon > 0$ can be chosen as close to zero as desired.

Convergence of AIS estimators

The results summarized above hold for general importance samplers. In an AIS framework, however, it is of specific interest to study the convergence of the estimators as the proposals are adapted. This issue is tackled in the classical article [40], where the estimators that result from aggregating weighted samples produced through several consecutive iterations are analyzed. Assuming that an AIS algorithm is run through J iterations, producing K samples per iteration for a total of JK samples overall (here we work with one proposal function per iteration), we construct the aggregated estimator of $I(f)$ as

$$\tilde{I}^{J\times K}(f) = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} f(\mathbf{x}_j^{(k)}) w_j^{(k)}}{\sum_{j=1}^{J}\sum_{k=1}^{K} w_j^{(k)}}. \tag{26}$$

In the setup of [40], the proposal functions $q_j(\mathbf{x})$ are selected from a parametric family $q(\mathbf{x};\boldsymbol{\theta})$, where $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_m]^\top \in \mathbb{R}^m$. The conditions to be satisfied by $q(\mathbf{x};\boldsymbol{\theta})$ are fairly general: $q(\mathbf{x};\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, the weight function $w = (\pi(\mathbf{x})/q(\mathbf{x};\boldsymbol{\theta}))$ is uniformly bounded (over the space of $\mathbf{x}$ and $\boldsymbol{\theta}$), and $q(\mathbf{x};\boldsymbol{\theta}) > 0$ whenever $\pi(\mathbf{x}) > 0$. In addition, it is assumed that there exists an optimal choice of the proposal function, of the form $q(\mathbf{x};\boldsymbol{\theta}_o)$, where $\boldsymbol{\theta}_o = E_{\tilde{\pi}}[\xi(\mathbf{x})]$ for some (possibly unknown) integrable function $\xi: \mathbb{R}^{d_x} \to \mathbb{R}^m$.

The latter is a regularity assumption: it implies that, if the weights are proper and $K \to \infty$, it is possible to approximate the target proposal $q(\mathbf{x};\boldsymbol{\theta}_o)$ as tightly as we wish. Under these assumptions, in [40] it is proved that

$$\lim_{J\times K\to\infty} \tilde{I}^{J\times K}(f) = I(f) \quad \text{a.s., and}$$
$$\lim_{J\times K\to\infty} \sqrt{JK}(\tilde{I}^{J\times K}(f) - I(f)) \overset{d}{=} \mathcal{N}(0,\sigma^2(f)), \tag{27}$$

where the limit variance $\sigma^2(f)$ is finite, and it depends on the test function and the normalization constant of $\tilde{\pi}$. Convergence of the first limit in (27) guarantees consistency, while the second expression is a CLT that shows that the asymptotic optimal error rate $O(1/JK)$ can be achieved without discarding any samples. Consistency of the aggregate estimator $\tilde{I}^{J\times K}(f)$ can be proved in a rather straightforward manner for most AIS schemes as long as the importance weights are proper at each iteration and the weight function remains bounded, even if an optimal or desired proposal $q(\mathbf{x};\boldsymbol{\theta}_o)$ does not exist (or simply changes from one iteration to the next).

AIS and high-dimensional target pdfs

The error bounds of (24) and (25) or the variances in the CLTs (22) and (27) depend on the dimension d_x of the target random vector $\mathbf{X}$, often in an intricate manner. Few analytical results on the effect of the dimension are available in the literature. In simplified scenarios, and through numerical studies, it has been shown that often the number of samples K has to be increased exponentially with d_x to attain a prescribed performance [41]. However, it has not been proved that this is necessarily the case, and some recent theoretical results actually suggest otherwise. In [42], the stability of the effective sample size (ESS), constructed as $ESS_j^K = (\Sigma_{k=1}^K w_j^{(k)})^2 \big/ \Sigma_{k=1}^K (w_j^{(k)})^2$, of a sequential MC sampler as the dimension increases, $d_x \to \infty$, is analyzed. The ESS, related to the variance of the weights, is commonly used to assess the numerical stability of the adaptive algorithms and detect the degeneracy phenomenon. In this AIS scheme, the target pdf $\tilde{\pi}(\mathbf{x})$ is approximated through a sequence of bridge densities $\pi_0(\mathbf{x}), \pi_1(\mathbf{x}), \ldots, \pi_j(\mathbf{x}), \ldots, \pi_J(\mathbf{x})$, where $\pi_0(\mathbf{x})$ is sufficiently easy to approximate via IS and $\pi_J(\mathbf{x}) = \tilde{\pi}(\mathbf{x})$. The intuition is that we can start approximating π_0 and, assuming $\pi_{j-1}(\mathbf{x})$ and $\pi_j(\mathbf{x})$ are similar enough, we can then move parsimoniously through the sequence of bridge pdfs until we obtain an approximation of $\tilde{\pi}(\mathbf{x}) = \pi_J(\mathbf{x})$. In this setup, the proposal functions $q_j(\mathbf{x})$ are devised as Markov kernels that jump from $\pi_{j-1}(\mathbf{x})$ to $\pi_j(\mathbf{x})$. In the specific scheme analyzed in [42], the bridge pdfs are constructed by tempering, i.e., selecting a sequence of positive real numbers $0 < \epsilon_0 < \epsilon_1 < \cdots < \epsilon_J = 1$ and then setting $\pi_j(\mathbf{x}) = \tilde{\pi}^{\epsilon_j}(\mathbf{x})$.

Under the strongly simplifying assumption of $\mathbf{X}$ being a vector of independent variables, i.e., $\tilde{\pi}(\mathbf{x}) = \prod_{i=1}^{d_x} \tilde{\pi}_i(x_i)$, but still assuming that the sample vector $\mathbf{x}_j^{(k)}$ is drawn jointly

(and not independently, entrywise) from the proposal $q_j(\mathbf{x})$, it is proved in [42] that $\lim_{d_x \to \infty} \mathrm{ESS}_j^K = C$ a.s., where C is a positive constant, even if the number of samples K is held constant. Moreover, this can be achieved when the number of bridge pdfs is $J = O(d_x)$. These results indicate that this particular AIS method remains numerically stable (i.e., the weights do not degenerate) as the dimension d_x becomes arbitrarily large; however, they are mainly of theoretical (rather than practical) interest because of the strong assumptions involved. Nevertheless, they suggest that AIS schemes may beat the curse of dimensionality in some scenarios if properly designed.

A comparison of the convergence properties of IS and MCMC methods

MCMC [43] and AIS methods are often competing techniques to tackle the same class of inference problems, hence a brief comparison of their theoretical properties is relevant. MCMC schemes generate a chain of correlated samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}$, using a suitable Markov kernel $\mathcal{K}(\mathbf{x}^{(k-1)}, \mathbf{x}^{(k)})$ to draw $\mathbf{x}^{(k)}$ conditional on $\mathbf{x}^{(k-1)}$. Different algorithms, e.g., the Gibbs sampler or the MH method [43], yield different kernels. In any case, $\mathcal{K}(\cdot, \cdot)$ is designed so as to guarantee, under mild assumptions, that $\lim_{k \to \infty} p_k = \tilde{\pi}$ a.s., where p_k denotes the pdf of the kth element of the chain, which generates $\mathbf{x}^{(k)}$, i.e., the generated sequence $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, has $\tilde{\pi}$ as a stationary pdf [7], [43], [44]. There are no known rates for the convergence of p_k toward $\tilde{\pi}$. However, it has been found that this rate can be very low in some scenarios. Moreover, it has to be taken into account that estimators constructed from an MCMC run of length K have the form

$$\tilde{I}_{MCMC}^K = \frac{1}{K - k_0} \sum_{k=k_0+1}^{K} f(\mathbf{x}^{(k)}), \tag{28}$$

where the first k_0 samples are discarded to allow for the convergence of p_k. While $E[\tilde{I}_{MCMC}^K(f)] \approx I(f)$, assuming $p_k \approx \tilde{\pi}$, the random variates $f(\mathbf{x}^{(k)})$ are correlated and, therefore, the analysis of $\mathrm{Var}(\tilde{I}_{MCMC}^K)$ is difficult. Again, it can be shown that $\tilde{I}_{MCMC}^K(f) \to I(f)$ a.s., but no error rates are available.

These double asymptotics inherent to MCMC [we need the chain to burn-in so that $p_k \to \tilde{\pi}$, then we need $K \to \infty$ for $\tilde{I}_{MCMC}^K(f) \to I(f)$] often make these algorithms slower and computationally less efficient than AIS schemes [32], [38]. Moreover, in problems where the normalizing constant $Z = \left(\int \pi(\mathbf{x}) d\mathbf{x}\right)^{-1}$ is of interest (e.g., for model validation or model selection), AIS is a natural solution, as it readily yields unbiased estimates $\hat{Z}_j^K = (1/K)\Sigma_{k=1}^K w^{(k)}$, $j = 1, \ldots, J$, while MCMC is often harder to apply [45]. There have been many recent attempts to devise algorithms that combine MCMC and AIS principles to take advantage of the strengths of both approaches [35], [46].

A pictorial comparison between IS and MCMC approaches is provided in Figure 5. In an MH-type sampler, a new state in the chain is proposed, and it is accepted or rejected with a suitable probability α. The number of repetitions of the same current state $\mathbf{x}^{(k)}$ plays the role of a weight in the estimator $\tilde{I}_{MCMC}^K(f)$. However, unlike in IS, given a sample $\mathbf{x}^{(k)}$, the weighting procedure is not provided by a deterministic function [e.g., by $\pi(\mathbf{x})/q(\mathbf{x})$] but instead is a result of a stochastic process defined by the acceptance MCMC tests performed at each iteration.

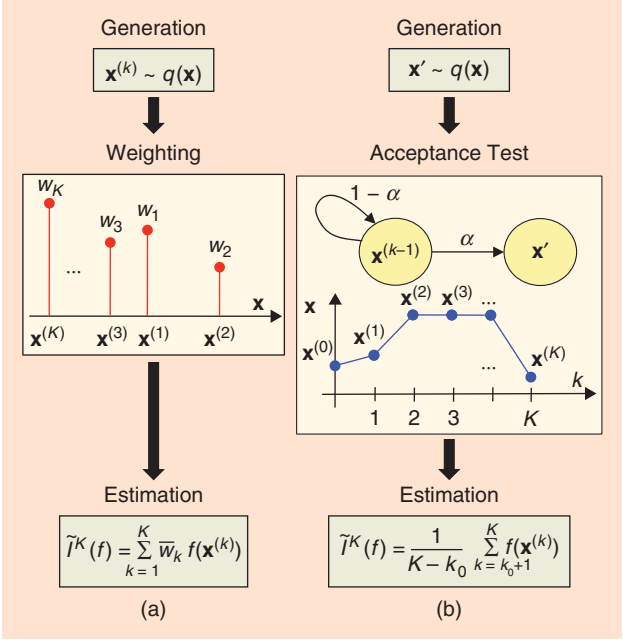


FIGURE 5. A graphical representation of IS and MCMC procedures to provide an estimator $\tilde{I}^K(f)$ of $I(f)$. More specifically, we have considered the MH type of MCMC algorithms, where a novel possible state $\mathbf{x}'$ is drawn from $q(x)$, and it is accepted, thus setting $\mathbf{x}^{(k)} = \mathbf{x}'$ with a suitable probability α. Otherwise, the next state of the chain is set equal to the previous one, i.e., $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$ with probability $1 - \alpha$. (a) The importance sampler and (b) the MH-type sampler.

Parallelization

IS methods are easily parallelizable, as the samples $\mathbf{x}^{(k)}$ are independent and, therefore, can be generated concurrently. In comparison, competing MCMC methods are much harder to parallelize, because the samples in a Markov chain are inherently sequential. With the availability of state-of-the-art multicore computers and graphics processing units (GPUs), this may be a key factor in favor of IS schemes. See [47] for a comparison of various MC schemes running on GPU systems.

In the specific case of AIS schemes, it is relatively straightforward to identify two stages in all of the presented algorithms. The first stage, which includes sampling and weighting, is a readily parallelizable task. This is the same as in standard IS, where each sample can (ideally) be generated and processed independently. The second stage, however, involves adaptation and, for some schemes, resampling. In this stage, it is necessary to process together all of the samples and weights, e.g., to calculate the parameters of the new proposals in schemes like AMIS or N-PMC, or even to run MCMC steps in the LAIS method. The adaptation step can be expected to be nonparallelizable, or parallelizable to a lesser extent, on standard computing devices.

Applications and challenges

While the range of applications of AIS algorithms is broad, it is worth discussing some particular fields where this methodology has either been applied with special success (compared to state-of-the-art techniques) or appears as a promising tool to tackle hard and long-standing problems.

The problems of detection and estimation in wireless sensor networks have been of great interest to the signal processing community for more than a decade. They involve scenarios where data related to a particular signal of interest are collected at various different sites of a network. Often, these observations can only be shared under tight constraints (due to scarce communication bandwidth, limited power, etc.), and estimation has to be performed with partial data or in a distributed fashion. One example of this class of problems, the localization of an object using signal-strength measurements, is presented in the "Localization Problem in a Wireless Sensor Network" section. A general challenge in this field is the design of schemes for the distributed implementation of AIS schemes with a minimal communication among the nodes of the network. Ideas based on the exchange of summary statistics have been explored, especially in the context of sequential IS (see, e.g., [48]), but efficient schemes (accurate yet affordable in terms of both communication and computation) are still needed.

The DM-PMC, LAIS, and GAPIS methods are particularly well suited to multimodal target distributions, which are often hard for conventional algorithms.

The fitting of Gaussian processes (GPs) for nonlinear regression problems is another example, which is explored in the section "Learning Hyperparameters for GP Regression Models." GPs have found a plethora of applications in problems where one needs to approximate smooth functions for which a parametric model is not available at all, and the complete function has to be learned from a discrete collection of data points [49]. While GPs are powerful models, their performance can be very sensitive to the fitting of a number of hyperparameters. The example in the aforementioned section shows that AIS can efficiently tackle this problem.

AIS has also shown advantages compared to state-of-the-art methods in performing inference for stochastic kinetic models (SKMs) [32]. SKMs are used in biochemistry or ecology to model complex interactions among populations of different species [50]. In ecology, SKMs yield a generalization of classical predator–prey models. In biochemistry, an SKM represents a system with n types of molecules (species) and k types of reactions. In both cases, it is of interest to track and predict the species populations, which evolve as a multidimensional continuous-time jump process, and estimate the rates that govern the dynamics. It has been shown [32] that AIS schemes (in this case, the N-PMC algorithm) can attain the same performance as state-of-the-art particle MCMC methods [51] with a fraction of the computational cost for modest SKMs. The accurate fitting of complex, high-dimensional SKMs is an open problem with outstanding real-world applications.

AIS techniques also enable consistent parameter estimation in α-stable distributions with very heavy tails [38]. α-stable distributions are often denoted as $S(\alpha,\beta,\gamma,\delta)$, where $0 < \alpha \leq 2$ determines the weight of the tails (the smaller the value of α, the heavier the tails), β is a skewness parameter, and $\gamma > 0$ and δ determine the scale and location. Except for particular cases, the associated pdfs can only be approximated numerically. Fast, classical methods for parameter estimation are known to work only for $\alpha \geq 0.5$ (i.e., with moderate tails). The results in [38], including an example with real data, show that AIS methods can overcome this limitation and open the door to address problems formerly intractable.

Finally, a challenging arena for the application of AIS methods includes a number of problems where very large-scale models are used and need to be fitted from (often scarce) data. This includes many large-scale systems used in geophysics, e.g., in oceanography [52], climate modeling [53] or cosmology [54]. In all of these cases, algorithms that attain a good tradeoff between computational complexity and accuracy of the resulting estimators are very much needed, and advanced AIS holds potential to be successfully applied.

Numerical examples

Localization problem in a wireless sensor network

We consider the problem of positioning a target in a wireless sensor network using range measurements [55]. We assume that the measurements of the sensors are contaminated by additive white Gaussian noise with different unknown powers. This situation is common in many practical scenarios where, even if the sensors are of the same manufacturer and model, the noise level can be different due to various factors. They include signal propagation conditions, manufacturing imperfections, and environmental conditions (e.g., humidity or temperature). Moreover, these conditions can change over time. Hence, in practice the central node of the network has to reestimate the noise powers (in addition to the target's position and possibly other parameters of the model) whenever a new block of observations is acquired.

More specifically, we denote the unknown target's position with the random vector $\boldsymbol{\Lambda} = [\Lambda_1, \Lambda_2]^\top$ and a specific realization of it as $\boldsymbol{\lambda}$. Let there be M sensors at locations $\mathbf{h}_m$, $m = 1, 2, \cdots, M$. The model for the observations is

$$y_{i,m} = 20\log(\|\boldsymbol{\lambda} - \mathbf{h}_m\|) + v_{i,m}, \quad m = 1, \ldots, M; \quad i = 1, 2, \ldots, N_o, \tag{29}$$

where $\|\cdot\|$ denotes the L_2 norm, $y_{i,m}$ is the ith observation of the Mth sensor, N_o is the number of observations of each of the sensors, and the $v_{i,m}$s are independent Gaussian random variables with pdfs $\mathcal{N}(v_{i,m}; 0, \gamma_m^2)$, $m = 1, \ldots, M$. We denote the vector of standard deviations as $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_M]$. We adopt

a uniform prior $\mathcal{U}(\mathcal{R}_\lambda)$ for the position $[\Lambda_1, \Lambda_2]^\top$, over a predefined support, and a uniform prior for γ_j, also over a preset range, $\mathcal{R}_\gamma$. Thus, the posterior pdf is

$$\tilde{\pi}(\lambda, \gamma | \mathbf{Y}) \propto \ell(\mathbf{y} | \lambda_1, \lambda_2, \gamma_1, \ldots, \gamma_M) \prod_{i=1}^{2} p(\lambda_i) \prod_{m=1}^{M} p(\gamma_m),$$
$$= \left[\prod_{i=1}^{N_o} \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\gamma_m^2}} \exp\left(-\frac{1}{2\gamma_m^2} (y_{i,m} - 20 \log(\| \boldsymbol{\lambda} - \mathbf{h}_m \|)^2 \right) \right] \times \mathbb{I}(\mathcal{R}_\lambda) \mathbb{I}(\mathcal{R}_\gamma), \quad (30)$$

where N_o is the number of observations, $y_{i,m}$ is the ith observation of the mth sensor, and $\mathbb{I}_c(\mathcal{S})$ is an indicator function that takes a value equal to one if $c \in \mathcal{S}$, and is equal to zero otherwise. Thus, in this problem $\mathbf{x} = [\boldsymbol{\lambda}^\top, \boldsymbol{\gamma}^\top]^\top$, and $d_x = M + 2$.

Our goal is to compute the minimum mean square error (MMSE) estimate, which corresponds to the expected value of the posterior $\tilde{\pi}(\boldsymbol{\lambda}, \boldsymbol{\gamma} | \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$, where the $\mathbf{y}_m$s are vectors whose elements are the measurements of the mth sensor. Because the MMSE estimate cannot be computed analytically, we applied several AIS methods to approximate it via MC quadrature. In particular, we worked with the standard PMC method [19], two different DM-PMC techniques [24], AMIS [21], and LAIS [23].

In our experiment, we had $M = 6$ sensors, and the locations of the sensors were at $\mathbf{h}_1 = [3, -8]^\top$, $\mathbf{h}_2 = [8, 10]^\top$, $\mathbf{h}_3 = [-4, -6]^\top$, $\mathbf{h}_4 = [-8, 1]^\top$, $\mathbf{h}_5 = [10, 0]^\top$, and $\mathbf{h}_6 = [0, 10]^\top$. In all of the cases, we employed Gaussian proposal densities, $q_{n,j}(\mathbf{x} | \boldsymbol{\mu}_{n,j}, \mathbf{C}_{n,j}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{n,j}, \mathbf{C}_{n,j})$ with $\boldsymbol{\mu}_{n,1} \sim \mathcal{U}([1,4]^{d_x})$ for $n = 1, \ldots, N$. The target was located at $\boldsymbol{\lambda} = [\lambda_1 = 2.5, \lambda_2 = 2.5]^\top$, and the vector of standard deviations was $\boldsymbol{\gamma} = [\gamma_1 = 1, \gamma_2 = 2, \gamma_3 = 1, \gamma_4 = 0.5, \gamma_5 = 3, \gamma_6 = 0.2]$. We generated $N_o = 20$ observations for each sensor according to the model given by (29). The uniform prior $\mathcal{U}(\mathcal{R}_\lambda)$ over the position $[\lambda_1, \lambda_2]^\top$ had a support $\mathcal{R}_\lambda = [-30 \times 30]^2$, and the uniform prior of the γ_is was $\mathcal{U}([0.01, 20])$. Thus, the overall prior of γ was $\mathcal{U}(\mathcal{R}_\gamma)$ with $\mathcal{R}_\gamma = [0.01, 20]^M$. Then, we obtained the measurement vectors $\mathbf{y}_1, \ldots, \mathbf{y}_M$, where $\mathbf{y}_i \in \mathbb{R}^{N_o}$. Note that, regarding the dimension of the observations, we have $d_y = N_o M = 120$.

For the PMC, the DM-PMCs and LAIS we set $\mathbf{C}_{n,j} = \mathbf{C}_n = \mathbf{C} = \sigma^2 \mathbf{I}$ with $\sigma = 1$. In AMIS, we have $N = 1$ and $\mathbf{C}_{n,j} = \mathbf{C}_j = \sigma_j^2 \mathbf{I}$, and we set $\sigma_1 \in \{1, 2\}$. In the adaptation layer of LAIS, to obtain $\{\boldsymbol{\mu}_{n,j}\}_{n=1}^N$, from the previous population $\{\boldsymbol{\mu}_{n,j-1}\}_{n=1}^N$, we employ parallel MH chains with a Gaussian random-walk proposal pdf, $\varphi_n(\boldsymbol{\mu}_{n,j} | \boldsymbol{\mu}_{n,j-1}, \sigma^2 \mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}_{n,j} | \boldsymbol{\mu}_{n,j-1}, \sigma^2 \mathbf{I})$ with $\sigma = 1$. Moreover, we also test the application of N independent parallel MH algorithms with the same Gaussian random-walk proposal pdf, $\varphi_n(\boldsymbol{\mu}_{n,t} | \boldsymbol{\mu}_{n,j-1}, \sigma^2 \mathbf{I})$, employed in the adaptation of LAIS.

We fix the total number of evaluations of the posterior density to $E = 10^4$, because this is usually the most costly step in MC algorithms. Let us recall that J denotes the total number of iterations and K the number of samples drawn from each proposal at each iteration. Moreover, we denote as S the total number of samples employed in the final IS estimator. In LAIS, the total number of evaluations of the target pdf is $E = NJ(K + 1)$, whereas $S = NJK$ (i.e., $E > S$ due to the use of the Markov adaptation process). For the rest of the methods, we have $E = S = NKJ$ (note that $N = 1$ in AMIS, while $K = 1$ in standard PMC and MH). Several combinations of N, J, and K are tested for the fixed $E = 10^4$ evaluations.

We computed the mean square error (MSE) of the different estimators obtained with respect to the ground truth, $\mathbf{x} = [\boldsymbol{\lambda}^\top, \boldsymbol{\gamma}^\top]^\top$. The results, averaged over 500 independent runs, are provided in Tables 7–12 (one table per technique) with the best and worst MSE values highlighted in boldface. In this particular experiment, with a unimodal posterior pdf and a good initialization $\mu_{n,1} \sim \mathcal{U}([1,5]^{d_x})$, the PMC techniques and the AMIS method provide the smallest MSE values. The standard PMC method seems to perform better if one uses a larger value of N and a smaller number of iterations J. In fact, the use of a small number of proposal pdfs can lead to catastrophic results in this case. The DM-PMC techniques substantially mitigate this problem, with

Table 7. The results of standard PMC [19] (localization example).

MSE	**25.12**	3.96	1.35	1.08	0.72	**0.61**	0.70
N	5	10	50	100	500	1,000	2,000
J	2000	1,000	200	100	20	10	5
E	$S = NJ = 10^4$						
Range	**MMSE = 0.61 —— Maximum MSE = 25.12**						

Table 8. The results of GR-DM-PMC [24] (localization example).

MSE	0.96	0.89	**0.75**	0.84	0.85	**1.47**	0.81	0.76	0.79	0.84	0.80	0.81
N	5	5	5	10	10	10	50	50	100	100	500	1,000
J	50	100	10	10	5	200	5	10	5	10	5	5
K	40	20	200	100	200	5	40	20	20	10	4	2
E	$S = NTM = 10^4$											
Range	**MMSE = 0.75 —— Maximum MSE = 1.47**											

Table 9. The results of LR-DM-PMC [24] (localization example).

MSE	1.14	1.52	**0.77**	**0.77**	0.79	**2.91**	1.01	1.24	1.26	1.44	1.32	1.49
N	5	5	5	10	10	10	50	50	100	100	500	1,000
J	50	100	10	10	5	200	5	10	5	10	5	5
K	40	20	200	100	200	5	40	20	20	10	4	2
E	$S = NTM = 10^4$											
Range	**MMSE = 0.77 —— Maximum MSE = 2.91**											

GR-DM-PMC showing a more robust behavior with respect to the parameter choice than LR-DM-PMC. AMIS provides very good results, although it shows some sensitivity with respect to the choice of the initial scale parameter, σ_1. Note that LAIS provides slightly worse results than AMIS but also shows less sensitivity with respect to the parameter choice and outperforms the performance of N independent parallel MH chains. Finally, Figure 6 shows the evolution of the estimators of AMIS $(J = 300, K = 200)$ and standard PMC $(N = 1{,}000, J = 100)$ as functions of the number of iterations, j, in one specific run.

Learning hyperparameters for GP regression models

GPs are a modern machine-learning approach to solving regression problems [56]. Given a covariance kernel function, learning its hyperparameters is the key to attain accurate performance. In this section, we test the different AIS schemes for estimating the hyperparameters of a GP regression model.

Let us assume that we have a set of observed data pairs, $\{y_i, \mathbf{z}_i\}_{i=1}^P$ with $y_i \in \mathbb{R}$ and $\mathbf{z}_i \in \mathbb{R}^L$, and let us denote the corresponding $P \times 1$ output vector as $\mathbf{y} = [y_1, \ldots, y_P]^\top$ and the $L \times P$ input matrix as $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_L]$. We address the problem of inferring the unknown function f that links the variables y and $\mathbf{z}$. Specifically, the assumed model is $y = f(\mathbf{z}) + e$, where $e \sim N(e; 0, \sigma^2)$ and $f(\mathbf{z})$ is a realization of a GP, $f(\mathbf{z}) \sim \mathcal{GP}(\mu(\mathbf{z}), \kappa(\mathbf{z}, \mathbf{r}))$ with $\mathbf{z}, \mathbf{r} \in \mathbb{R}^L$, $\mu(\mathbf{z}) = 0$, and the kernel function has the form

$$\kappa(\mathbf{z}, \mathbf{r}) = \exp\left(-\sum_{\ell=1}^{L} \frac{(z_\ell - r_\ell)^2}{2\alpha^2}\right). \tag{31}$$

[We point out that $f(\cdot)$ in this section has nothing to do with the test function used previously in the article.] Given these assumptions, the vector $\mathbf{f} = [f(\mathbf{z}_1), \ldots, f(\mathbf{z}_P)]^\top$ is distributed as $p(\mathbf{f} \mid \mathbf{Z}, \alpha, \kappa) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$, where $\mathbf{0}$ is a $P \times 1$ null vector, and $[\mathbf{K}]_{ij} := \kappa(\mathbf{z}_i, \mathbf{z}_j)$ for all $i, j = 1, \ldots, P$ is a $P \times P$ matrix. Therefore, $d_x = 2$, and the vector containing the hyperparameters of the model is $\mathbf{x} = [x_1 = \alpha, x_2 = \sigma] \in \mathbb{R}^2$, where α is the hyperparameter of the kernel function in (31), and σ is the standard deviation of the observation noise. In this experiment, we focus on the marginal posterior density of the hyperparameters [56], $\tilde{\pi}(\mathbf{x} \mid \mathbf{y}, \mathbf{Z}, \kappa) \propto p(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \kappa) p(\mathbf{x})$, which can be evaluated analytically, but we cannot compute integrals involving it. Considering a uniform prior $p(\mathbf{x})$ over $[0.01, 20]^2$, and because $p(\mathbf{y} \mid \mathbf{x}, \mathbf{Z}, \kappa) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I})$, we have

$$\log[\pi(\mathbf{x} \mid \mathbf{y}, \mathbf{Z}, \kappa)] = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} - \frac{1}{2}\log[\det(\mathbf{K} + \sigma^2 \mathbf{I})], \tag{32}$$

Table 10. The results of AMIS [21] (localization example).

MSE ($\sigma_0 = 1$)	0.80	**0.72**	0.75	0.76	0.88	1.29
MSE ($\sigma_0 = 2$)	1.53	1.48	1.42	1.29	1.48	**1.71**
N	1					
J	200	100	50	20	10	5
K	50	100	200	500	1,000	2,000
E	$S = TM = 10^4$					
Range	**MMSE = 0.72 —— Maximum MSE = 1.71**					

Table 11. The results of LAIS [23] (localization example).

MSE	**1.91**	1.52	1.14	1.11	1.10	**1.06**	1.29	1.25	1.26	1.30	1.41
N	1	2	5	5	10	10	100	100	100	200	10^3
J	$5{\cdot}10^3$	500	250	500	250	500	10	25	50	25	5
K	1	9	7	3	3	1	9	3	1	1	1
S	$5{\cdot}10^3$	$9{\cdot}10^3$	8,750	7,500	7,500	$5{\cdot}10^3$	$9{\cdot}10^3$	7,500	$5{\cdot}10^3$	$5{\cdot}10^3$	$5{\cdot}10^3$
E	$S + NT = NT(M + 1) = 10^4$										
Range	**MMSE = 1.06 —— Maximum MSE = 1.91**										

where $\mathbf{K}$ depends on α [56]. Because the moments of this marginal posterior cannot be computed analytically, we use again MC integration with different AIS methods to approximate the mmse estimator, $\hat{\boldsymbol{x}} = [\hat{\alpha}, \hat{\sigma}]$, which corresponds to the expected value of $\mathbf{X}$ with respect to $\tilde{\pi}(\mathbf{x} \mid \mathbf{y}, \mathbf{Z}, \kappa)$.

For this experiment, we generated $P = 200$ pairs of data, $\{y_j, \mathbf{z}_j\}_{j=1}^{P}$, according to the previous GP model with $\alpha = 3$, $\sigma = 10$, $L = 1$, and $z_j \sim \mathcal{U}([0, 10])$. Fixing the generated data, we then computed the true value of the MMSE, $\hat{\mathbf{x}} = [\hat{\alpha}, \hat{\sigma}] \approx [3.5200, 9.2811]$, using an exhaustive and costly grid search approximation, to compare the different AIS techniques. The corresponding posterior pdf is given in Figure 7(a).

Table 12. The results of independent MH parallel chains (localization example).

MSE	1.42	**1.31**	1.44	2.32	2.73	**3.21**	3.18	3.15
N	1	5	10	50	100	500	1,000	2,000
J	10^4	$2 \cdot 10^3$	10^3	200	100	20	10	5
E	$S = NT = 10^4$							
MSE range	**MMSE = 1.31 —— Maximum MSE = 3.21**							

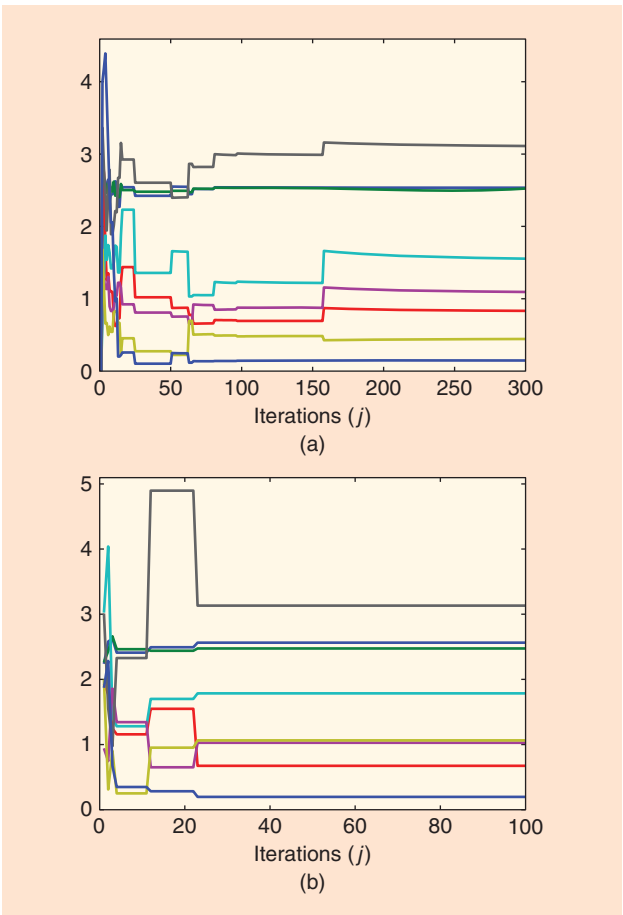


FIGURE 6. The evolution of (a) the estimators of AMIS ($T = 300$, $M = 200$) and (b) standard PMC ($N = 1{,}000$, $T = 100$) as functions of the number of iterations, j, in one specific run. The true values of the parameters are $x_1 = 2.5$ (green), $x_2 = 2.5$ (blue), $x_3 = 1$ (yellow), $x_4 = 2$ (cyan), $x_5 = 1$ (magenta), $x_6 = 0.5$ (red), $x_7 = 3$ (black), and $x_8 = 0.2$ (violet).

We compared the standard PMC method [19], the LR-DM-PMC technique [24], the AMIS [21], and the LAIS [23] algorithms. Again, for all of them we considered Gaussian proposal densities, $q_{n,j}(\mathbf{x} \mid \boldsymbol{\mu}_{n,j}, \mathbf{C}_{n,j}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_{n,j}, \mathbf{C}_{n,j})$ with $\boldsymbol{\mu}_{n,1} \sim \mathcal{U}([1,4]^2)$ for $n = 1, \ldots, N$. Note that, unlike in the previous experiment, the true value of $\mathbf{x}$ does not belong to the initialization region $[1, 4]^2$. For PMC, LR-DM-PMC, and LAIS we set $\mathbf{C}_{n,j} = \mathbf{C}_n = \mathbf{C} = \sigma^2 \mathbf{I}$ with $\sigma = 2$. For AMIS, we had $N = 1$ and $\mathbf{C}_{n,j} = \mathbf{C}_j = \sigma_j^2 \mathbf{I}$, and we set $\sigma_1 = 2$. In the adaptation layer of LAIS, to obtain $\{\boldsymbol{\mu}_{n,j}\}_{n=1}^{N}$ from the previous population $\{\boldsymbol{\mu}_{n,j-1}\}_{n=1}^{N}$, we employed parallel MH chains with a Gaussian random-walk proposal pdf, $\varphi_n(\boldsymbol{\mu}_{n,j} \mid \boldsymbol{\mu}_{n,j-1}, \sigma^2 \mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}_{n,j} \mid \boldsymbol{\mu}_{n,j-1}, \sigma^2 \mathbf{I})$ with $\sigma = 2$. Once more, we fixed the total number of evaluations of the posterior pdf to $E = 10^4$, and we tested the algorithms considering different combinations of the parameters.

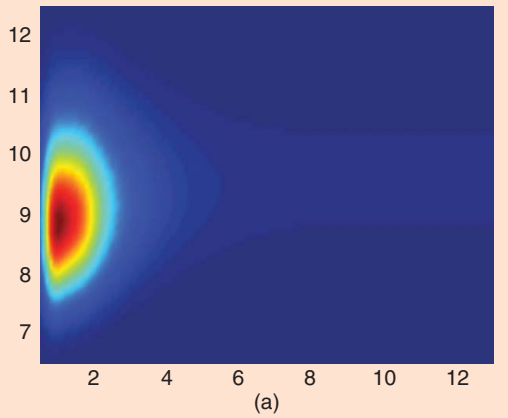
The results, in terms of MSE in the estimation of $\mathbf{x}$, are given in Tables 13–16. They were averaged over 500 independent runs. In this numerical experiment, LAIS and LR-DM-PMC provided smaller MSEs because they discover and explore faster the tail of the posterior distribution with respect to the other techniques. The adaptation of the location parameters produced in one specific run by LAIS ($N = 5$ and $T = 100$) is shown in Figure 7(b).

Concluding remarks and outlook

In signal processing, an important task is making inference from data about model parameters or models in general. From a Bayesian point of view, ideally, this inference is made from posterior distributions of the unknowns. For complex models, it is very difficult to find these posteriors. In such cases, one resorts to approximations in the sense that one generates samples that are drawn from the posterior distributions. A tool that helps practitioners to get such samples is MCMC sampling. As has already been pointed out, the MCMC algorithms and the growth of computing power have invigorated the Bayesian methodology in the last 25 years to the point that today we use it to solve most intricate problems.

In this article, we have argued that practitioners of signal processing should be aware of another option for solving inference problems by way of drawing samples from distributions. It is based on a methodology known as AIS. AIS methods have the subtle ability to learn the pdfs that produce better samples for constructing posteriors and that eventually allow for a more accurate inference. The learning is accomplished in iterations where the samples from previous iterations serve to find better proposal pdfs.

AIS is often simpler to implement than MCMC sampling. Besides simplicity, AIS has other advantages over MCMC sampling, including that it does not produce correlated samples, there is no such thing as burn-in period, and AIS is easier for parallelization. We also have a better understanding



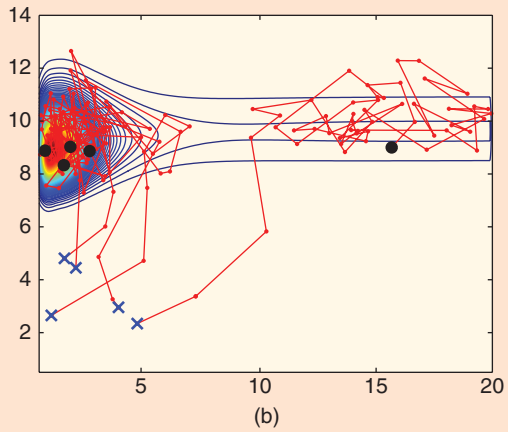


FIGURE 7. (a) The posterior density $\pi(\mathbf{x}|\mathbf{y},\mathbf{Z},\kappa)$. (b) The evolution of the location parameters $\mu_{n,t}$ in one specific run of LAIS with $N = 5$ and $T = 100$ (jointly with the contour plot of the posterior pdf). The starting points are shown with x marks, whereas the final locations are depicted with circles.

Table 13. The results of standard PMC [19] (GP example).

MSE	**0.44**	0.87	1.01	0.88	0.86	0.95	**1.15**
N	5	50	100	200	500	1,000	2,000
T	2,000	200	100	50	20	10	5
E	$S = NT = 10^4$						
Range	**MMSE = 0.44 —— Maximum MSE = 1.15**						

Table 14. The results of LR-DM-PMC [24] (GP example).

MSE	0.41	0.39	0.16	0.09	**0.04**	0.23	0.07	**0.46**
N	5	5	5	50	50	100	100	1,000
T	10	20	40	10	20	10	20	5
M	200	100	50	20	10	10	5	2
E	$S = NTM = 10^4$							
Range	**MMSE = 0.04 —— Maximum MSE = 0.46**							

of the rates of convergence of AIS methods than those of MCMC sampling. A pitfall of IS methods is the possibility of using proposal pdfs with thinner tails than those of the target distribution, which can easily ruin any estimate from the generated data and the computed weights.

The most important open problem of AIS, as we have already alluded, is the development of AIS methods that can work accurately in high-dimensional spaces. As the dimension of $\mathbf{x}$ increases, the complexity of finding good proposal pdfs explodes (curse of dimensionality). One approach for resolving this problem is to work with compartmentalized spaces of the unknowns and accept that we will not have approximations of the full joint posterior but instead a number of marginalized posteriors.

In the years to come, we expect that AIS methods will find increased use within the signal processing community.

Another way of addressing high dimensionality is by particle flows. This approach has been of interest in particle filtering, where samples drawn from the prior distribution are migrated to the posterior distribution of the unknowns by solving partial differential equations [57]. Even though the problems we solve with AIS are different from those addressed by particle filtering, there is enough common ground between the two methodologies to investigate the application of particle flows to AIS. How can the underlying principles of particle flows be exploited in AIS?

In recent years, stochastic optimization methods have seen a resurgence. One reason for this is that there are many problems that can be formulated as optimization problems, in which the minimized objective function is a sum of many loss functions. IS is one of a number of MC sampling-based methods for stochastic optimization. It can improve the convergence rate of the optimization and reduce the stochastic variance of the result [58]. The use of AIS for optimization raises various challenging questions, including convergence to optimal solutions and optimal values.

A specific application of stochastic optimization is in stochastic variational Bayesian methods. These methods can be applied to complex probabilistic models and large data sets with a vast range of applications in machine learning. Recently, a synthesis between variational inference and MCMC sampling for variational approximation has been proposed [59]. It was claimed that a fast posterior approximation through the maximization of an explicit objective was accomplished. Furthermore, the proposed method offered tradeoffs between computation and accuracy. Clearly, AIS is a natural candidate to be applied in the same setting with the possibility of performing even better than MCMC sampling.

Finally, in the years to come, we expect that AIS methods will find increased use within the signal processing community. Much of the research in this area will be driven by novel applications and by models with expanded complexity. There will be new applications that may even include use of

Table 15. The results of AMIS [21] (GP example).

MSE	1.32	**1.35**	1.26	1.27	**1.23**
N	1				
T	200	100	50	20	10
M	50	100	200	500	1,000
E	$S = TM = 10^4$				
Range	**MMSE = 1.23 —— Maximum MSE = 1.35**				

Table 16. The results of LAIS [23] (GP example).

MSE	**1.04**	0.46	0.21	0.09	**0.03**	0.31	0.65
N	1	5	10	50	100	500	1,000
T	5,000	1,000	500	100	50	10	5
M	1						
E	$NT(M+1) = 10^4$						
Range	**MMSE = 0.03 —— Maximum MSE = 1.04**						

AIS in deep learning for computing the weights of the hidden layers. The addressed problems will not only require estimating unknown quantities but also finding the best models from a set of predefined models or finding the best model in nonparametric Bayesian settings where the number of models is not set a priori.

Acknowledgments

We gratefully acknowledge the National Science Foundation under awards CCF-1617986 (Mónica F. Bugallo) and CCF-1618999 (Petar M. Djurić); Ministerio de Economía y Competitividad of Spain under TEC2015-69868-C2-1-R ADVENTURE and the Office of Naval Research Global under award N62909-15-1-2011 (Joaquín Míguez); the European Research Council (ERC) through the ERC Consolidator Grant SEDAL ERC-2014-CoG 647423 (Luca Martino); MINECO of Spain under TEC2015-64835-C3-3-R MIMOD-PLC project, Ministerio de Educación, Cultura y Deporte of Spain under CAS15/00350 grant, and Universidad Politécnica de Madrid through a mobility grant for a short visit to Stony Brook University (David Luengo) and MINECO of Spain through Red de Excelencia KERMES TEC2016-81900-REDT (David Luengo and Luca Martino).

Authors

Mónica F. Bugallo (monica.bugallo@stonybrook.edu) received her B.S., M.S., and Ph.D. degrees in computer science and engineering from the University of A Coruña, Spain. She is an associate professor of electrical and computer engineering and the faculty director of the Women in Science and Engineering program at Stony Brook University, New York. Her research interests are in the field of statistical signal processing, with emphasis on the theory of Monte Carlo methods and their application to different disciplines including biomedicine, sensor networks, and finance. She has authored and coauthored two book chapters and more than 150 journal papers and refereed conference articles. She is a Senior Member of the IEEE.

Víctor Elvira (victor.elvira@imt-lille-douai.fr) received his B.S., M.S., and Ph.D. degrees in electrical engineering from the Universidad de Cantabria, Spain, in 2007, 2008, and 2011, respectively. Currently, he is with IMT Lille Douai, Villeneuve d'Ascq, France, and Université de Lille and CRIStAL Laboratory (UMR 9189), Villeneuve d'Ascq, France. He is a Member of the IEEE.

Luca Martino (luca.martino@uv.es) received his M.S. degree in electronic engineering from the Politecnico di Milano, Italy, and his Ph.D. degree in statistical signal processing in 2011 from the Universidad Carlos III de Madrid, Spain. Currently, he is with the Image Processing Laboratory, Universitat de València, Spain.

David Luengo (david.luengo@upm.es) received his M.S. and Ph.D. degrees in electrical engineering from the Universidad de Cantabria, Spain, in 1998 and 2006, respectively. From 2003 to 2011, he was an assistant professor with the Universidad Carlos III de Madrid, Spain. Since 2011, he has been an associate professor at the Universidad Politécnica de Madrid. His research interests include statistical signal processing, Monte Carlo methods, and multitask learning. He is a Member of the IEEE.

Joaquín Míguez (joaquin.miguez@uc3m.es) received his M.S. and Ph.D. degrees in computer engineering from the University of A Coruña, Spain, in 1997 and 2000, respectively. Currently, he is with the department of signal theory and communications, Universidad Carlos III de Madrid. His research interests are in the fields of applied probability, statistical signal processing, Bayesian analysis, dynamical systems and the theory and applications of Monte Carlo methods. He is a corecipient of the 2007 IEEE Signal Processing Magazine Best Paper Award.

Petar M. Djurić (petar.djuric@stonybrook.edu) received his B.S. and M.S. degrees in electrical engineering from the University of Belgrade, Serbia, and his Ph.D. degree in electrical engineering from the University of Rhode Island. He is currently a distinguished professor in the Department of Electrical and Computer Engineering at Stony Brook University, New York. His research has been in signal and information processing, with an emphasis on Monte Carlo-based methods, signal processing over networks and applications in wireless sensor networks, and radio-frequency identification. He received the IEEE Signal Processing Magazine Best Paper Award in 2007 and the EURASIP Technical Achievement Award in 2012. He is a Fellow of the IEEE and EURASIP.

References

[1] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, 2005.

[2] Z. Chen, F. Xia, T. Huang, F. Bu, and H. Wang, "A localization method for the Internet of Things," *J. Supercomput.*, pp. 1–18, 2013.

[3] C. Phillips, J. Mattout, M. D. Rugg, P. Maquet, and K. J. Friston, "An empirical Bayesian solution to the source reconstruction problem in eeg," *NeuroImage*, vol. 24, no. 4, pp. 997–1011, 2005.

[4] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1595–1608, 2016.

[5] I. Shmulevich and E. R. Dougherty, *Genomic Signal Processing*. Princeton, NJ: Princeton Univ. Press, 2014.

[6] N. Metropolis, "The beginning of the Monte Carlo method," *Los Alamos Sci.*, no. 15, pp. 125–130, 1987.

[7] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2004.

[8] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[9] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[10] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: CRC, 1995.

[11] H. Kahn, "Random sampling (Monte Carlo) techniques in neutron attenuation problems," *Nucleonics*, vol. 6, no. 5, pp. 27–33, 1950.

[12] T. Hesterberg, "Weighted average importance sampling and defensive mixture distributions," *Technometrics*, vol. 37, pp. 185–194, 1995.

[13] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 1994.

[14] N. Metropolis and S. Ulam, "The Monte Carlo method," *J. Amer. Statistical Assoc.*, vol. 44, pp. 335–341, 1949.

[15] G. Rubino and B. Tuffin, *Rare Event Simulation Using Monte Carlo Methods*. Hoboken, NJ: Wiley, 2009.

[16] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovskii, Eds. London, U.K.: Oxford Univ. Press, 2009, pp. 656–704.

[17] E. Veach and L. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," in *Proc. 22nd Annu. Conf. Computer Graphics and Interactive Techniques* (*SIGGRAPH*), 1995, pp. 419–428

[18] A. Owen and Y. Zhou, "Safe and effective importance sampling," *J. Amer. Statistical Assoc.*, vol. 95, no. 449, pp. 135–143, 2000.

[19] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert, "Population Monte Carlo," *J. Computational Graphical Statist.*, vol. 13, no. 4, pp. 907–929, 2004.

[20] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert, "Adaptive importance sampling in general mixture classes," *Statist. Computing*, vol. 18, pp. 447–459, 2008.

[21] J. M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian J. Statist.*, vol. 39, no. 4, pp. 798–812, Dec. 2012.

[22] L. Martino, V. Elvira, D. Luengo, and J. Corander, "An adaptive population importance sampler: Learning from the uncertainty," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4422–4437, 2015.

[23] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statist. Computing*, vol. 27, no. 3, pp. 599–623, 2017.

[24] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Improving population Monte Carlo: Alternative weighting and resampling schemes," *Signal Process.*, vol. 131, no. 12, pp. 77–91, 2017.

[25] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," submitted for publication.

[26] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Efficient multiple importance sampling estimators," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1757–1761, 2015.

[27] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Heretical multiple importance sampling," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1474–1478, 2016.

[28] M. F. Bugallo, L. Martino, and J. Corander, "Adaptive importance sampling in signal processing," *Dig. Signal Process.*, vol. 47, pp. 36–49, 2015.

[29] T. Li, M. Bolic, and P. M. Djurić, "Resampling methods for particle filtering: Classification, implementation, and strategies," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 70–86, 2015.

[30] G. R. Douc, J.-M. Marin, and C. Robert, "Convergence of adaptive mixtures of importance sampling schemes," *Ann. Statist.*, vol. 35, pp. 420–448, 2007.

[31] G. R. Douc, J.-M. Marin, and C. Robert, "Minimum variance importance sampling via population Monte Carlo," *ESAIM: Probability Statist.*, vol. 11, pp. 427–447, 2007.

[32] E. Koblents and J. Míguez, "A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models," *Statist. Computing*, vol. 25, no. 2, pp. 407–425, 2015.

[33] R. J. Steele, A. E. Raftery, and M. J. Emond, "Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS)," *J. Computational Graphical Statist.*, vol. 15, pp. 712–734, 1996.

[34] V. Elvira, L. Martino, D. Luengo, and J. Corander, "A gradient adaptive population importance sampler," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, pp. 4075–4079.

[35] Z. I. Botev, P. L'Ecuyer, and B. Tuffin, "Markov chain importance sampling with applications to rare event probability estimation," *Statist. Computing*, vol. 23, no. 2, pp. 271–285, 2013.

[36] P. Del Moral, *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. New York: Springer-Verlag, 2004.

[37] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, vol. 24, pp. 1317–1399, 1989.

[38] E. Koblents, J. Miguez, M. A. Rodriguez, and A. M. Schmidt, "A nonlinear population Monte Carlo scheme for the Bayesian estimation of parameters of α-stable distributions," *Computational Statist. Data Anal.*, vol. 95, pp. 57–74, Mar. 2016.

[39] D. Williams, *Probability with Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[40] M.-S. Oh and J. O. Breger, "Adaptive importance sampling in Monte Carlo integration," *J. Statistical Computation Simulation*, vol. 41, nos. 3–4, pp. 143–168, 1992.

[41] T. Bengtsson, P. Bickel, and B. Li, "Curse of dimensionality revisited: Collapse of particle filter in very large scale systems," *Probability Statistics*, vol. 2, pp. 316–334, 2008.

[42] A. Beskos, D. Crisan, and A. Jasra, "On the stability of sequential Monte Carlo methods in high dimensions," *Ann. Appl. Probability*, vol. 24, no. 4, pp. 1396–1445, 2014.

[43] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. New York: Taylor & Francis, 1995.

[44] S. Asmussen and P. W. Glynn, "A new proof of convergence of MCMC via the ergodic theorem," *Statist. Probability Lett.*, vol. 81, no. 10, pp. 1482–1485, 2011.

[45] C. P. Robert, *The Bayesian Choice*. New York: Springer-Verlag, 2007.

[46] L. Martino, V. Elvira, D. Luengo, and J. Corander, "MCMC-driven adaptive multiple importance sampling," in *Interdisciplinary Bayesian Statistics*. New York: Springer-Verlag, 2015, pp. 97–109.

[47] A. Lee, C. Yau, M. B. Giles, C. C. Doucet, and A. Holmes, "On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods," *J. Computational Graphical Statist.*, vol. 19, no. 4, pp. 769–789, 2010.

[48] O. Hlinka, O. Slučiak, F. Hlawatsch, P. Djurić, and M. Rupp, "Likelihood consensus and its application to distributed particle filtering," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4334–4349, 2012.

[49] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, 2005.

[50] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood, "Bayesian inference for a discretely observed stochastic kinetic model," *Statist. Computing*, vol. 18, no. 2, pp. 125–135, 2008.

[51] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. Roy. Statistical Soc. B*, vol. 72, no. 3, pp. 269–342, 2010.

[52] C. K. Wikle, R. F. Milliff, D. Nychka, and L. M. Berliner, "Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds," *J. Amer. Statistical Assoc.*, vol. 96, no. 454, pp. 382–397, 2001.

[53] J. Rougier, "Probabilistic inference for future climate using an ensemble of climate model evaluations," *Climatic Change*, vol. 81, no. 3, pp. 247–264, 2007.

[54] A. Lewis, "Efficient sampling of fast and slow cosmological parameters," *Phys. Rev. D, Part. Fields*, vol. 87, no. 10, p. 103,529, 2013.

[55] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[56] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[57] Y. Li and M. Coates, "Particle filtering with invertible particle flow," submitted for publication.

[58] T. Homem-de Mello and G. Bayraksan, "Monte Carlo sampling-based methods for stochastic optimization," *Surveys Operations Res. Manage. Sci.*, vol. 19, no. 1, pp. 56–85, 2014.

[59] T. Salimans, D. Kingma, and M. Welling, "Markov chain Monte Carlo and variational inference: Bridging the gap," in *Proc. 32nd Int. Conf. Machine Learning* (*ICML-15*), 2015, pp. 1218–1226.

SP

Yubin Deng, Chen Change Loy, and Xiaoou Tang

Image Aesthetic Assessment

An experimental survey

This article reviews recent computer vision techniques used in the assessment of image aesthetic quality. Image aesthetic assessment aims at computationally distinguishing high-quality from low-quality photos based on photographic rules, typically in the form of binary classification or quality scoring. A variety of approaches has been proposed in the literature to try to solve this challenging problem. In this article, we summarize these approaches based on visual feature types (hand-crafted features and deep features) and evaluation criteria (data set characteristics and evaluation metrics). The main contributions and novelties of the reviewed approaches are highlighted and discussed. In addition, following the emergence of deep-learning techniques, we systematically evaluate recent deep-learning settings that are useful for developing a robust deep model for aesthetic scoring.

Experiments are conducted using simple yet solid baselines that are competitive with the current state of the art. Moreover, we discuss the possibility of manipulating the aesthetics of images through computational approaches. We hope that this article might serve as a comprehensive reference for future research on the study of image aesthetic assessment.

Aesthetic Assessment Through Computer Vision

The aesthetic quality of an image is judged by commonly established photographic rules, which can be affected by numerous factors, including the different uses of lighting [1], contrast [2], and image composition [3] [see Figure 1(a)]. These human judgments, given in an aesthetic evaluation setting, are the result of human aesthetic experience, i.e., the interaction between emotional–valuation, sensory–motor, and meaning–knowledge neural systems, as demonstrated in a systematic neuroscience study by Chatterjee et al. [4]. From the beginning of psychological aesthetics studies by Fechner [5] to modern neuroaesthetics, researchers have argued that there is a certain connection between human aesthetic experience and the sensation caused by visual stimuli, regardless of source, culture, and experience [6], which is supported by activations in specific regions of the visual cortex [7]–[10]. For example, humans' general reward circuitry produces pleasure when they look at beautiful objects [11], and the subsequent aesthetic judgment consists of the appraisal of the valence of such

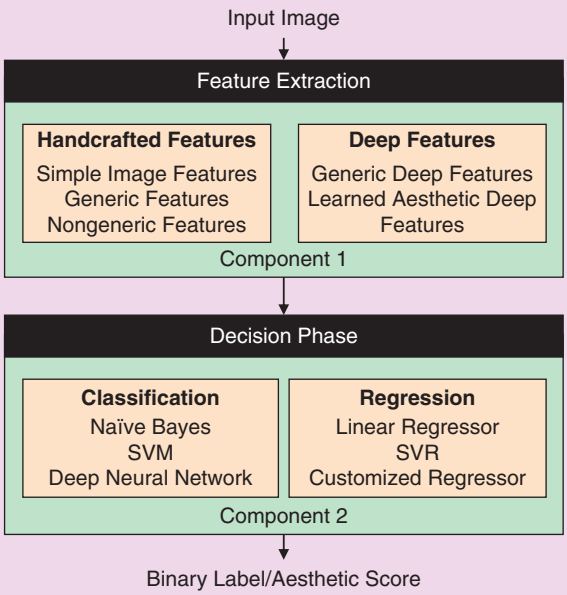
Digital Object Identifier 10.1109/MSP.2017.2696576
Date of publication: 11 July 2017

©GRAPHICSTOCK

perceived objects [8]–[10], [12]. These activations in the visual cortex can be attributed to the processing of various early, intermediate, and late visual features of the stimuli, including orientation, shape, color grouping, and categorization [13]–[16]. Artists intentionally incorporate such features to facilitate desired perceptual and emotional effects in viewers, forming a set of guidelines as they create artworks to induce desired responses in the nervous systems of perceivers [16], [17]. And modern photographers, to make their work appealing to as large an audience as possible, now also resort to certain well-established photographic rules [18], [19] when they capture images.

As the volume of visual data available online grows at an exponential rate, the capability of automatically distinguishing high-quality images from low-quality ones is in increasing demand in real-world image searching and retrieving applications. When a person enters a particular keyword in an image search

(a)



(b)

FIGURE 1. (a) Some high-quality images following well-established photographic rules (top row: color harmony; middle row: single salient object and low depth of field; bottom row: black-and-white portraits with decent lighting contrast). (b) A typical flow of image aesthetic assessment systems. SVM: support vector machine; SVR: support vector regressor.

engine, it is expected that the system will return professional photographs instead of random snapshots. For example, when a user enters the words "mountain scenery," the person will expect to see colorful, pleasing mountain views or well-captured mountain peaks instead of gray or blurry mountain snapshots.

The design of these intelligent systems can potentially be facilitated by insights from neuroscience studies, which show that human aesthetic experience is a kind of information processing that includes five stages: perception, implicit memory integration, explicit classification of content and style, cognitive mastering, and evaluation, which together ultimately produce aesthetic judgment and aesthetic emotion [12], [13]. However, it is nontrivial to computationally model this process. Challenges in the task of judging the quality of an image include 1) computationally modeling the intertwined photographic rules, 2) knowing the aesthetic differences in images from different image genres (e.g., close-shot object, profile, scenery, and night scenes), 3) knowing the type of techniques used in photo capturing (e.g., high-dynamic range, black and white, and depth of field), and 4) obtaining a large amount of human-annotated data for robust testing.

To address these challenges, computer vision researchers typically cast this problem as a classification or regression problem. Early studies started with distinguishing typical snapshots from professional photographs by trying to model the well-established photographic rules using low-level features [20]–[22]. These systems typically involve a training set and a testing set consisting of both high-quality and low-quality images. The system robustness is judged by the model performance on the testing set using a specified metric, such as accuracy. These rule-based approaches are intuitive, as they try to explicitly model the criteria that humans use in evaluating the aesthetic quality of an image. However, more recent studies [23]–[26] have shown that using a data-driven approach is more effective, as the amount of training data available grows from a couple of hundred images to millions. Besides, transfer learning from source tasks with sufficient amounts of data to a target task with relatively fewer training data is also proven feasible, with many successful attempts showing promising results through deep-learning methods [27] with network fine-tuning, where image aesthetics are implicitly learned in a data-driven manner.

As summarized in Figure 1(b), the majority of the aforementioned computer vision approaches for image aesthetic assessment can be categorized based on image representations (e.g., handcrafted features and learned features) and classifiers/regressors training (e.g., support vector machine [SVM] and neural network learning approaches). To the best of our knowledge, no up-to-date survey covers the state-of-the-art methodologies involved in image aesthetic assessment. The last review was published in 2011 by Joshi et al. [28], and no deep learning-based methods were covered. Some reviews on image-quality assessment have been published [29], [30]. In those efforts, image-quality metrics regarding the differences between a noise-tempered sample and the original high-quality image were proposed, including but not limited to mean squared error, structural similarity index (SSIM) [31], and visual information fidelity (VIF) [32]. Nevertheless, their main

focus was on distinguishing noisy images from clean ones in terms of a different quality measure rather than artistic/photographic aesthetics.

In this article, we contribute a thorough overview of the field of image aesthetic assessment. Meanwhile, we also cover the basics of deep-learning methodologies. Specifically, as different data sets exist and evaluation criteria vary in the image aesthetics literature, we do not aim to directly compare the system performance of all of the reviewed works; instead, we point out in the survey their main contributions and novelties in model designs, and give potential insights for future directions in this field of study. In addition, following the recent emergence of deep-learning techniques and the effectiveness of the data-driven approach in learning better image representation, we systematically evaluate different techniques that could facilitate the learning of a robust deep classifier for aesthetic scoring. Our study covers topics such as data preparation, fine-tuning strategies, and multicolumn deep architectures, which we believe to be useful for researchers working in this domain.

In particular, we summarize useful insights on how to alleviate the potential problem of data distribution bias in a binary classification setting and show the effectiveness of rejecting false-positive predictions using our proposed convolutional neural network (CNN) baselines, as revealed by the balanced accuracy metric. We also review the most commonly used publicly available image aesthetic assessment data sets for this problem and draw connections between image aesthetic assessment and image aesthetic manipulation, including image enhancement, computational photography, and automatic image cropping.

Background

The deep neural network

The deep neural network belongs to the family of deep-learning methods that are tasked to learn feature representation in a data-driven approach. While shallow models (e.g., SVM and boosting) showed success in earlier studies concerning relatively smaller amounts of data, they require highly engineered feature designs in solving machine-learning problems. Common architectures in deep neural networks consist of a stack of parameterized individual modules that we call *layers*, such as the convolution layer and the fully connected layer. The architecture design of stacking layers on top of layers is inspired by the hierarchy in the human visual cortex ventral pathway, offering different levels of abstraction for the learned representation in each layer. Information propagation among layers in feed-forward deep neural networks typically follows a sequential pattern. A forward operation $F(\cdot)$ is defined respectively in each layer to propagate the input $\boldsymbol{x}$ it receives and produces an output $\boldsymbol{y}$ to the next layer. For example, the forward operation in a fully connected layer with learnable weights $\mathbf{W}$ can be written as

$$y = F(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x} = \sum w_{ij} \cdot x_i. \tag{1}$$

This is typically followed by a nonlinear function, such as sigmoid

$$z = \frac{1}{1 + \exp(-y)} \tag{2}$$

or the rectified linear unit $z = \max(0, y)$, which acts as the activation function and produces the net activation output z.

To learn the weights $\mathbf{W}$ in a data-driven manner, we need to have the feedback information that reports the current performance of the network. Essentially, we are trying to tune the knobs $\mathbf{W}$ to achieve a learning objective. For example, given an objective t for the input $\boldsymbol{x}$, we want to minimize the squared error between the net output z and t by defining a loss function L:

$$L = \frac{1}{2} \| z - t \|^2. \tag{3}$$

To propagate this feedback information to the weights, we define the backward operation for each layer using gradient backpropagation [33]. We hope to get the direction $\Delta\mathbf{W}$ to update the weights $\mathbf{W}$ to better suit the training objective (i.e., to minimize L): $\mathbf{W} \leftarrow \mathbf{W} - \eta\Delta\mathbf{W}$, where η is the learning rate. In our example, $\Delta\mathbf{W}$ can be easily derived based on the chain rule:

$$\begin{aligned} \Delta\mathbf{W} &= \frac{\partial L}{\partial \mathbf{W}} \\ &= \frac{\partial L}{\partial z} \frac{\partial z}{\partial y} \frac{\partial y}{\partial \mathbf{W}} \\ &= (z - t) \cdot \frac{\exp(-y)}{(\exp(-y) + 1)^2} \cdot \boldsymbol{x}. \end{aligned} \tag{4}$$

In practice, researchers resort to batch stochastic gradient descent or more advanced learning procedures that compute more stable gradients, as averaged from a batch of training examples $\{(\boldsymbol{x}_i, t_i) \mid \boldsymbol{x}_i \in X\}$ to train deeper and deeper neural networks with continually increasing numbers of layers. We refer readers to [27] for an in-depth overview of additional deep-learning methodologies.

Image-quality metrics

Image-quality metrics are defined in an attempt to quantitatively measure the objective quality of an image. This is typically used in image restoration applications (superresolution [34], deblurring [35], and deartifacting [36]), where we have a default high-quality reference image for comparison. However, these quality metrics are not designed to measure the subjective nature of human-perceived aesthetic quality (see examples in Figure 2). Directly applying these objective quality metrics to our domain of image aesthetic assessment may produce misleading results, as can be seen from the measured values in Figure 2(b). Interest in developing more robust metrics has increased in the research community, as a means to assess the more subjective quality of image aesthetics.

A typical pipeline

Most existing image-quality assessment methods take a supervised learning approach. A typical pipeline assumes a set of training data $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$, from which a function $f: g(X) \rightarrow Y$ is learned, where $g(\boldsymbol{x}_i)$ denotes the feature representation of image $\boldsymbol{x}_i$. The label $\boldsymbol{y}_i$ is either $\{0, 1\}$ for binary classification (when f is a classifier) or a continuous score range for regression (when f is a regressor). Following this formulation, a pipeline can be broken into two main components, as shown in Figure 1(b), i.e., a feature extraction component and a decision component.

Feature extraction

The first component of an image aesthetics assessment system aims at extracting robust feature representations describing the aesthetic aspect of an image. Such features are assumed to model the photographic/artistic aspect of images to distinguish images of different qualities. Numerous efforts have been made to design features that are robust enough for the intertwined aesthetic rules. The majority of feature types can be classified into handcrafted features and deep features. Conventional approaches [20], [21], [37]–[49] typically adopt handcrafted features to computationally model the photographic rules (e.g., lighting and contrast), global image layout (the rule of thirds), and typical objects (e.g., human profiles, animals, and plants) in images. In more recent work, generic deep features [50], [51] and learned deep features [23]–[25], [52]–[59] exhibit stronger representation power for this task.

Decision phase

The second component of an image aesthetics assessment system provides the ability to perform classification or regression for the given aesthetic task. The naïve Bayes classifier, SVM, boosting, and deep classifier are typically used for binary classification of high-quality and low-quality images, whereas regressors like support vector regressors (SVRs) are used in ranking or scoring images based on their aesthetic quality.

FIGURE 2. Quality measurements by peak signal-to-noise ratio (PSNR), SSIM [31], and VIF [32] (a higher measurement is better, typically made against a referencing ground-truth high-quality image). Although these are good indicators for measuring the quality of images in image restoration applications, such as the images in (a), they do not reflect human-perceived aesthetic values, as shown by the measurements for the building images in (b).

Data sets

The assessment of image aesthetic quality assumes a standard training set and testing set containing both high-quality and low-quality image examples, as previously mentioned. Judging the ground-truth aesthetic quality of a given image is, however, a subjective task. As such, it is inherently challenging to obtain a large amount of such annotated data. Most of the earlier papers [21], [38], [39] on image aesthetic assessment collect a small amount of private image data. These data sets typically contain from a few hundred to a few thousand images, with binary labels or aesthetic scoring for each image. Yet such data sets where the model performance is evaluated are not publicly available. Much research effort has later been made to contribute publicly available image aesthetic data sets of larger scale for more standardized evaluation of model performance. In the following, we introduce those data sets that are most frequently used in performance benchmarking for image aesthetic assessment.

The Photo.net data set and the DPChallenge data set are introduced in [28] and [60], respectively. These two data sets can be considered the earliest attempts to construct large-scale image databases for image aesthetic assessment. The Photo.net data set contains 20,278 images, with at least ten score ratings per image. The ratings range from zero to seven, with seven assigned to the most aesthetically pleasing photos. Typically, images uploaded to Photo.net are rated as somewhat pleasing, with the peak of the global mean score skewing to the right in the distribution [28]. The more challenging DPChallenge data set contains diverse ratings. The DPChallenge data set contains 16,509 images in total, and was later replaced by the Aesthetic Visual Analysis (AVA) data set, where a significantly larger number of images derived from DPChallenge.com are collected and annotated.

The Chinese University of Hong Kong-PhotoQuality (CUHK-PQ) data set is introduced in [45] and [61]. It contains 17,690 images collected from DPChallenge.com and amateur photographers. All of the images are given binary aesthetic labels and grouped into seven scene categories, i.e., animals, plants, static, architecture, landscape, humans, and night. The standard training and testing set from this data set are random partitions of a 50–50 split or a fivefold cross-validation partition, where the overall ratio of the total number of positive examples and that of the negative examples is around 1:3. Sample images are shown in Figure 3.

The AVA data set [49] contains ~250,000 images in total. These images are obtained from DPChallenge.com and labeled by aesthetic scores. Specifically, each image receives $78 \sim 549$ votes of scores ranging from one to ten. The average score of an image is commonly taken to be its ground-truth label. As such, it contains more challenging examples, as images that lie within the center score range could be aesthetically ambiguous [Figure 4(a)]. For the task of binary aesthetic quality classification, images with an average score higher than a threshold of $5 + \sigma$ are treated as positive examples, and images with an average score lower than $5 - \sigma$ are treated as negative ones. Additionally, the AVA data set contains 14 style attributes and more than 60 category attributes for a subset of images. There are two typical training and testing splits from this data set, i.e., 1) a large-scale standardized partition with ~230,000 training images and ~20,000 testing images using a hard threshold of $\sigma = 0$, and 2) an easier partition modeling that of CUHK-PQ by taking those images whose score ranking is at the top 10% and the bottom 10%, resulting in ~25,000 images for training and ~25,000 images for testing. The ratio of the total number of positive examples to that of the negative examples is around 12:5.

Apart from these two standard benchmarks, more recent research also introduces new data sets that take into consideration the data-balancing issue. The Image Aesthetic Data Set (IAD) introduced in [55] contains 1.5 million images derived from DPChallenge and Photo.net. Similar to AVA, images in the IAD data set are scored by annotators. Positive examples are selected from those images with a mean score larger than a threshold. All IAD images are used for model training, and the model performance is evaluated on AVA in [55]. The ratio of the number of positive examples to that of the negative

FIGURE 3. Some sample images in the CUHK-PQ data set [45]. (a) Distinctive differences can be visually observed between the high-quality (grouped in the green-framed box) and low-quality images (grouped in the red-framed box). (b) The number of images in the CUHK-PQ data set.

examples is around 1.07:1. The Aesthetic and Attributes Database (AADB) [25] also contains a balanced distribution of professional and consumer photos, with a total of 10,000 images. Eleven aesthetic attributes and annotators' IDs are provided. A standard partition with 8,500 images for training, 500 images for validation, and 1,000 images for testing is proposed [25].

The trend toward creating data sets of even larger volume and higher diversity is essential for boosting the research progress in this field of study. To date, the AVA data set serves as a canonical benchmark for performance evaluation of image aesthetic assessment, as it is the first large-scale data set with detailed annotation. Still, the distribution of positive and negative examples in the data set also plays a role in the effectiveness of trained models, as false-positive predictions are as harmful as having a low recall rate in image retrieval and searching applications. In the following, we review major attempts in the literature to build systems for the challenging task of image aesthetic assessment.

The distribution of positive and negative examples in the data set also plays a role in the effectiveness of trained models.

Conventional approaches with handcrafted features

The conventional option for image quality assessment is to hand-design good feature extractors, which requires a considerable amount of engineering skill and domain expertise. Next we review a variety of approaches that exploit hand-engineered features.

Simple image features

Global features are first explored by researchers to model the aesthetic aspect of images. The works by Datta et al. [21] and Ke et al. [37] are among the first to cast aesthetic understanding of images into a binary classification problem. Datta et al. [21] combine low-level and high-level features that are typically used for image retrieval and train an SVM classifier for binary classification of images in terms of aesthetic quality. Ke et al. [37] propose global edge distribution, color distribution, hue count, and low-level contrast and brightness indicators to represent an image; then they train a naïve Bayes classifier based on such features. An even earlier attempt by Tong et al. [20] adopts boosting to combine global low-level simple features (blurriness, contrast, colorfulness, and saliency) to classify professional photographs and ordinary snapshots.

All of these pioneering works present the very first attempts to computationally model the global aesthetic aspect of images using handcrafted features. Even in a recent work, Ayd in

FIGURE 4. Some sample images in the AVA data set [49]. (a) Images in the green-framed box are labeled with a mean score of >5. Images in the red-framed box are labeled with a mean score of <5. The image groups on the right are ambiguous, with a somewhat neutral scoring around five. (b) The number of images in the AVA data set.

et al. [62] construct image aesthetic attributes by sharpness, depth, clarity, tone, and colorfulness. An overall aesthetics rating score is heuristically computed based on these five attributes. Improving upon these global features, later studies adopt global saliency to estimate aesthetic attention distribution. Sun et al. [38] make use of a global saliency map to estimate visual attention distribution to describe an image, and they train a regressor to output the quality score of an image based on the rate-of-focused-attention region in the saliency map. You et al. [39] derive similar attention features based on a global saliency map and incorporate a temporal activity feature for video quality assessment.

Regional image features [40]–[42] later prove to be effective in complementing the global features. Luo et al. [40] extract regional clarity contrast, lighting, simplicity, composition geometry, and color harmony features based on the subject region of an image. Wong et al. [63] compute exposure, sharpness, and texture features on salient regions and global images, as well as features depicting the subject–background relationship of an image. Nishiyama et al. [41] extract bags-of-color patterns from local image regions with a grid-sampling technique. While [40], [41], and [63] adopt the SVM classifier, Lo et al. [42] build a statistical modeling system with coupled spatial relations after extracting color and texture features from images, where a likelihood evaluation is used for aesthetic quality prediction. These methods focus on modeling image aesthetics from local image regions that are potentially most attractive to humans.

The rule of thirds, low depth of field, and opposing colors are the common techniques for composing a good image where the salient object is made outstanding.

Image composition features

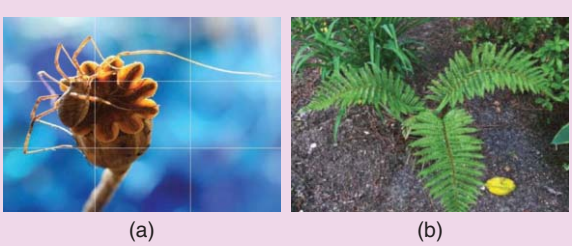
Image composition in a photograph typically relates to the presence and position of a salient object. The rule of thirds, low depth of field, and opposing colors are the common techniques for composing a good image where the salient object is made outstanding (see Figure 5). To model such aesthetic aspects, Bhattacharya et al. [43], [64] propose compositional features using relative foreground position and a visual weight ratio to model the relations between foreground objects and the background scene; then an SVR is trained. Wu et al. [65] propose the use of Gabor filter responses to estimate the position of the main object in images, and then extract low-level hue, saturation, value (HSV)-color features from global and central image regions. These features are fed to a soft-SVM classifier with sigmoidal softening to distinguish images of ambiguous quality. Dhar et al. [44] cast high-level features into describable attributes of composition, content, and sky illumination and combine low-level features to train an SVM classifier. Lo et al. [66] propose the combination of layout composition, edge composition features with an HSV color palette, HSV counts, and global features (textures, blur, dark channel, and contrasts). SVM is used as the classifier.

The representative work by Tang et al. [45] gives a comprehensive analysis of the fusion of global features and regional features. Specifically, image composition is estimated by global hue composition and scene composition, and multiple types of regional features extracted from subject areas are proposed, such as dark channel feature, clarity contrast, lighting contrast, composition geometry of the subject region, spatial complexity and human-based features. An SVM classifier is trained on each of the features for comparison, and the final model performance is substantially enhanced by combining all of the proposed features. It is shown that regional features can effectively complement global features in modeling the image aesthetics.

A more recent approach by image composition features is proposed by Zhang et al. [67], where image descriptors that characterize local and global structural aesthetics from multiple visual channels are designed. The spatial structure of the image local regions is modeled using graphlets, and they are connected based on atomic region adjacency. To describe such atomic regions, visual features from multiple visual channels [such as color moment, histogram of oriented gradients (HOG), and saliency histogram] are used. The global spatial layout of the photo is also embedded into graphlets using a Grassmann manifold. The importance of the two kinds of graphlet descriptors is dynamically adjusted, capturing the spatial composition of an image from multiple visual channels. The final aesthetic prediction of an image is generated by a probabilistic model using the postembedding graphlets.

General-purpose features

Yeh et al. [46] make use of scale-invariant feature transform (SIFT) descriptors and propose relative features by matching a query photo to photos in a gallery group. General-purpose imagery features like bag of visual (BOV) words [68] and Fisher vector (FV) [69] are explored in [47]–[49]. Specifically, SIFT and color descriptors are used as the local descriptors upon which a Gaussian mixture model (GMM) is trained. The statistics up to the second order of this GMM distribution are



(a) (b)

FIGURE 5. (a) An image composition with low depth of field, a single salient object, and the rule of thirds [49]. (b) An image of low aesthetic quality [45].

then encoded using the BOV words or FV. Spatial pyramid is also adopted, and the per-region encoded FVs are concatenated as the final image representation. These methods ([47]–[49]) represent an attempt to implicitly model photographic rules by encoding them in generic content-based features, which is competitive with or even outperforms simple handcrafted features.

Task-specific features

Task-specific features is a term that refers to features in image aesthetic assessment that are optimized for a specific category of photos, which can be efficient when the use-case or task scenario is fixed or known beforehand. Explicit information (such as human facial characteristics, geometry tag, scene information, or intrinsic character component properties) is exploited based on the different task nature.

Li et al. [70] propose a regression model that targets only consumer photos with faces. Face-related social features (such as facial expression features, facial pose features, and relative facial position features) and perceptual features (facial distribution symmetry, facial composition, and pose consistency) are specifically designed for measuring the quality of images with faces, and it is shown in [70] that for this task they complement conventional handcrafted features (brightness contrast, color correlation, clarity contrast, and background color simplicity). Support vector regression is used to produce aesthetic scores for images.

Lienhard et al. [71] study particular facial features for evaluating the aesthetic quality of headshot images. To design features for face/headshots, the input image is divided into subregions (the eyes, mouth, global face, and entire image regions). Low-level features (sharpness, illumination, contrast, dark channel, and hue and saturation in the HSV color space) are computed from each region. These pixel-level features assume the human way of perceiving a facial image and hence can reasonably model the headshot images. SVM with Gaussian kernel is used as the classifier.

Su et al. [72] propose bag of aesthetics-preserving features for scenic/landscape photographs. Specifically, an image is decomposed into $n \times n$ spatial grids; then low-level features in HSV-color space as well as local binary patterns, HOG, and saliency features are extracted from each patch. The final feature is generated by a predefined patch-wise operation to exploit the landscape composition geometry. AdaBoost is used as the classifier. These features aim at modeling only landscape images and may be limited in their representation power in general image aesthetic assessment.

Yin et al. [73] build a scene-dependent aesthetic model by incorporating the geographic location information with GIST descriptors and spatial layout of saliency features for scene aesthetic classification (such as bridges, mountains, and beaches). SVM is used as the classifier. The geographic location information is used to link a target scene image to relevant photos taken within the same geocontext; then these relevant photos are used as the training partition to the SVM. The authors' proposed model requires input images with geographic tags and is also limited to scenic photos. For scene images without geo-context information, SVM trained with images from the same scene category is used.

Sun et al. [74] design a set of low-level features for aesthetic evaluation of Chinese calligraphy. They target the handwritten Chinese character on a plain white background; hence, conventional color information is not useful in this task. Global shape features, extracted based on standard calligraphic rules, are introduced to represent a character. In particular, the authors consider alignment and stability, distribution of white space, stroke gaps, and a set of component layout features while modeling the aesthetics of handwritten characters. A backpropagation neural network is trained as the regressor to produce an aesthetic score for each given input.

Deep-learning approaches

The powerful feature representation learned from a large amount of data has shown an ever-improving performance in the tasks of recognition, localization, retrieval, and tracking, surpassing the capability of conventional handcrafted features [75]. Since the work by Krizhevsky et al. [75], where CNNs are adopted for image classification, a great degree of interest has arisen in learning robust image representations through deep-learning approaches. Recent works in the literature of image aesthetic assessment using deep-learning approaches to learn image representations can be broken down into two major schemes: 1) adopting generic deep features learned from other tasks and training a new classifier for image aesthetic assessment and 2) learning aesthetic deep features and training a classifier directly from image aesthetics data.

Generic deep features

A straightforward approach to employing deep-learning aims is to adopt generic deep features learned from other tasks and train a new classifier on the aesthetic classification task. Dong et al. [50] propose adopting the generic features from the penultimate layer output of AlexNet [75] with spatial pyramid pooling. Specifically, the 4,096 (fc7) $\times$ 6 (SpatialPyramid) = 24,576 - dimensional feature is extracted as the generic representation for images; then an SVM classifier is trained for binary aesthetic classification. Lv et al. [51] also adopt the normalized 4,096-dimension fc7 output of AlexNet [75] for feature representation. They propose to learn the relative ordering relationship of images of different aesthetic quality. They use SVM rank [76] to train a ranking model for image pairs of $\{I_{\text{HighQuality}}, I_{\text{LowQuality}}\}$.

Learned aesthetic deep features

Features learned with single-column CNNs

Peng et al. [52] propose to train CNNs of AlexNet-like architecture for eight different abstract tasks (emotion classification, artist classification, artistic style classification, aesthetic classification, fashion style classification, architectural style classification, memorability prediction, and interestingness

prediction). (Figure 6 illustrates a typical single-column CNN.) In particular, the last layer of the CNN for aesthetic classification is modified to output two-dimensional softmax probabilities. This CNN is trained from scratch using aesthetic data, and the penultimate layer (fc7) output is used as the feature representation. To further analyze the effectiveness of the features learned from other tasks, Peng et al. analyze different pretraining and fine-tuning strategies and evaluate the performance of different combinations of the concatenated fc7 features from the eight CNNs.

Wang et al. [53] propose a CNN that is modified from the AlexNet architecture. Specifically, the conv_5 layer of AlexNet is replaced by a group of seven convolutional layers (with respect to different scene categories), which are stacked in a parallel manner with mean pooling before feeding to the fully connected layers, i.e., $\{\text{conv}_5^{1-\text{animal}}, \text{conv}_5^{2-\text{architecture}}, \text{conv}_5^{3-\text{human}}, \text{conv}_5^{4-\text{landscape}}, \text{conv}_5^{5-\text{night}}, \text{conv}_5^{6-\text{plant}}, \text{conv}_5^{7-\text{static}}\}$. The fully connected layers fc6 and fc7 are modified to output 512 feature maps instead of 4,096 for more efficient parameter learning. The 1,000-class softmax output is changed to two-class softmax (fc8) for binary classification. The advantage of this CNN using such a group of seven parallel convolutional layers is to exploit the aesthetic aspects in each of the seven scene categories. During pretraining, a set of images belonging to one of the scene categories is used for each of the $\text{conv}_5^i (i \in \{1,...,7\})$ layers. Then the weights learned through this stage are transferred back to the conv_5^i in the proposed parallel architecture, with the weights from conv_1 to conv_4 reused from AlexNet in the fully connected layer randomly reinitialized. Subsequently, the CNN is further fine-tuned end to end. Upon convergence, the network produces a strong response in the conv_5^i layer feature map when the input image is of category $i \in \{1,...,7\}$. This shows the potential in exploiting image category information when learning the aesthetic presentation.

Tian et al. [54] train a CNN with four convolution layers and two fully connected layers to learn aesthetic features from the data. The output size of the two fully connected layers is set to 16 instead of 4,096 as in AlexNet. The authors propose that such a 16-dimension representation is sufficient to model only the top 10% and bottom 10% of the aesthetic data, which are relatively easy to classify compared to the full data. Based on this efficient feature representation learned from the CNN, the authors propose a query-dependent aesthetic model as the classifier. Specifically, for each query image, a query-dependent training set is retrieved based on predefined rules (visual similarity, image tags association, or a combination of both). Subsequently, an SVM is trained on this retrieved training set. It shows that the features learned from the aesthetic data outperform the generic deep features learned in the ImageNet task.

The deep multipatch aggregation (DMA)-net is proposed in [24], where information from multiple image patches is extracted by a single-column CNN that contains four convolution layers and three fully connected layers, with the last layer outputting a softmax probability. Each randomly sampled image patch is fed into this CNN. To combine multiple feature outputs from the sampled patches of one input image, a statistical aggregation structure is designed to aggregate the features from the orderless sampled image patches by multiple poolings (minimum, maximum, median, and averaging). An alternative aggregation structure is also designed based on sorting. The final feature representation effectively encodes the image based on regional image information.

Features learned from multicolumn CNNs

The Rating Pictorial Aesthetics using Deep Learning (RAPID) model by Lu et al. [23], [55] can be considered to be the first attempt to train CNNs with aesthetic data. They use an AlexNet-like architecture where the last fully connected layer is set to output two-dimensional probability for aesthetic binary classification. Both global image and local image patches are considered in their network input design, and the best model is obtained by stacking a global-column and a local-column CNN to form a double-column CNN, where the feature representation (the penultimate layers' fc7 output) from each column is concatenated before the fc8 layer (classification layer). (Figure 7 shows a typical multicolumn CNN.) Standard stochastic gradient descent is used to train the network with softmax loss. Moreover, the authors further boost the performance of the network by incorporating image style information using a style-column or semantic-column CNN. Then the style-column CNN is used as the third input column, forming a three-column CNN with style/semantic information. Such a multicolumn CNN exploits the data from both the global and local image aspects.

Mai et al. [26] propose stacking five columns of Visual Geometry Group (VGG)-based networks using an adaptive spatial pooling layer. The adaptive spatial pooling layer is

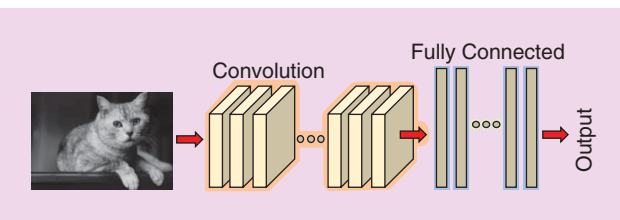


FIGURE 6. The architecture of a typical single-column CNN [49].

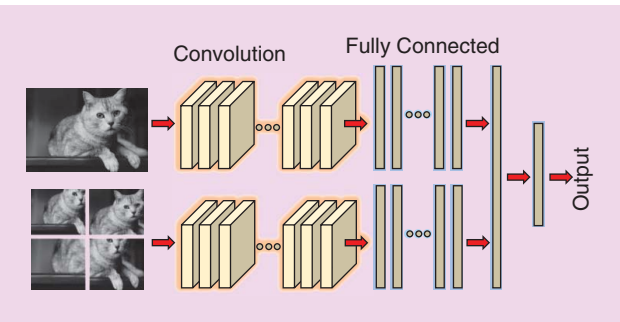


FIGURE 7. A typical multicolumn CNN (a two-column architecture is shown as an example) [49].

designed to allow arbitrary-sized images as input; specifically, it pools a fixed-length output, given different receptive field sizes, after the last convolution layer. By varying the kernel size of the adaptive pooling layer, each subnetwork effectively encodes multiscale image information. Moreover, to potentially exploit the aesthetic aspect of different image categories, a scene categorization CNN outputs a scene category posterior for each input image. Then a final scene-aware aggregation layer processes such aesthetic features (category posterior and multiscale VGG features) and outputs the final classification label. The design of this multicolumn network has the advantage of being able to exploit the multiscale composition of an image in each subcolumn by adaptive pooling, yet the multiscale VGG features may contain redundant or overlapping information, which could potentially lead to network overfitting.

Wang et al. [56] propose a multicolumn CNN model called *brain-inspired deep networks* (*BDN*) that shares similar structures with RAPID. In RAPID, a style attribute prediction CNN is trained to predict 14 style attributes for input images. This attribute CNN is treated as one additional CNN column, which is then added to the parallel input pathways of a global image column and a local patch column. In BDN, 14 different style CNNs are pretrained, and they are parallel cascaded and used as the input to a final CNN for rating distribution prediction, where the aesthetic quality score of an image is subsequently inferred. The BDN model can be considered as an extended version of RAPID that exploits each of the aesthetic attributes using learned CNN features, hence enlarging the parameter space and learning capability of the overall network.

Zhang et al. [57] propose a two-column CNN for learning aesthetic feature representation. The first column (CNN_1) takes image patches as input, and the second column (CNN_2) takes a global image as input. Instead of randomly sampling image patches, given an input image, a weakly supervised learning algorithm is used to project a set of D textual attributes learned from image tags to highly responsive image regions. Such image regions in images are then fed to the input of CNN_1. This CNN_1 contains four convolution layers and one fully connected layer (fc_5) at the bottom. Then a parallel group of D output branches ($\text{fc}_6^i, i \in \{1,2,...,D\}$) modeling each of the D textual attributes are connected on top. The size of the feature maps of each of the fc_6^i is of 128 dimensions. A similar CNN_2 takes a globally warped image as input, producing one more 128-dimension feature vector from fc_6. Hence, the final concatenated feature learned in this manner is $128 \times (D+1)$ dimensional. A probabilistic model containing four layers is trained for aesthetic quality classification.

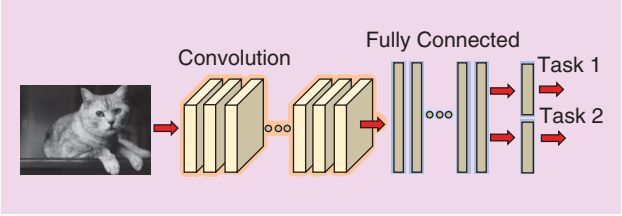


FIGURE 8. A typical multitask CNN consists of a main task (task 1) and multiple auxiliary tasks, only one of which is shown here (task 2) [49].

Kong et al. [25] propose learning aesthetic features assisted by the pair-wise ranking of image pairs as well as the image attribute and content information. Specifically, a Siamese architecture that takes image pairs as input is adopted, where the two base networks of the Siamese architecture adopt the AlexNet configurations (the 1,000-class classification layer fc8 from the AlexNet is removed). In the first stage, the base network is pretrained by fine-tuning from aesthetic data using the Euclidean loss regression layer instead of the softmax classification layer. After that, the Siamese network ranks the loss for every sampled image pair. Upon convergence, the fine-tuned base network is used as a preliminary feature extractor.

In the second stage, an attribute prediction branch is added to the base network to predict image attribute information. Then the base network continues to be fine-tuned in a multitask manner by combining the rating regression Euclidean loss, attribute classification loss, and ranking loss.

In the third stage, yet another content classification branch is added to the base network to predict a predefined set of category labels. Upon convergence, the softmax output of the content category prediction is used as a weighting vector for weighting the scores produced by each feature branch (the aesthetic branch, attribute branch, and content branch).

In the final stage, the base network and all of the added output branches are fine-tuned jointly, with the content classification branch frozen. Effectively, such aesthetic features are learned by considering both the attribute and category content information, and the final network produces image scores for each given image.

Features learned with multitask CNNs

Kao et al. [58] propose three category-specific CNN architectures: one for object, one for scene, and one for texture. The scene CNN takes a warped global image as input. It has five convolution layers and three fully connected layers, with the last fully connected layer producing a two-dimensional softmax classification. The object CNN takes both the warped global image and the detected salient region as input. It is a two-column CNN combining global composition and salient information. The texture CNN takes 16 randomly cropped patches as input. Category information is predicted using a three-class SVM classifier before feeding images to a category-specific CNN. To alleviate the use of the SVM classifier, an alternative architecture with a warped global image as input is trained with a multitask approach, where the main task is aesthetic classification and the auxiliary task is scene category classification. (A typical multitask CNN is illustrated in Figure 8.)

Kao et al. [59] propose learning image aesthetics in a multitask manner. Specifically, AlexNet is used as the base network. Then the 1,000-class fc8 layer is replaced by a two-class

aesthetic prediction layer and a 29-class semantic prediction layer. The loss balance between the aesthetic prediction task and the semantic prediction task is determined empirically. Moreover, another branch containing two fully connected layers for aesthetic prediction is added to the second convolution layer ($conv_2$ of AlexNet). By linking an added gradient flow from the aesthetic task directly to the convolutional layers, one expects to learn better low-level convolutional features. This strategy shares a similar spirit with the deeply supervised net [77].

Evaluation criteria and existing results

Different metrics for performance evaluation of image aesthetic assessment models are used across the literature: classification accuracy [20], [21], [23]–[25], [40], [43], [47], [49], [50], [52]–[59], [63]–[65], [71], [73] reports the proportion of correctly classified results; precision-and-recall (PR) curve [37], [40], [41], [44], [66] considers the degree of relevance of the retrieved items and the retrieval rate of relevant items, which is also widely adopted in image search or retrieval applications; Euclidean distance or residual sum-of-squares error between the ground-truth score and aesthetic ratings [38], [70], [71], [74] and correlation ranking [25], [39], [46] are used for performance evaluation in score regression frameworks; receiver-operating characteristic (ROC) curve [42], [48], [66], [71], [72] and area under the curve [45], [61], [66] concerns the performance of binary classifiers when the discrimination threshold is varied; mean average precision [23], [24], [51], [55] is the average precision (AP) across multiple queries, which is usually used to summarize the PR curve for the given set of samples. These are among the typical metrics for evaluating model effectiveness for image aesthetic assessment (see Table 1 for a summary). Subjective evaluation by conducting human surveys is also seen in [62], where human evaluators are asked to give subjective aesthetic attribute ratings.

We find that it is not feasible to directly compare all methods, as different data sets and evaluation criteria are used across the literature. To this end, we try to summarize, respectively, the released results reported on the two standard data sets, namely the CUHK-PQ (Table 2) and AVA data sets (Table 3), and to present the results on other data sets in Table 4. To date, the AVA data set (standard partition) is considered to be the most challenging by the majority of the reviewed work.

The overall accuracy metric appears to be the most popular metric. It can be written as

$$\text{Overall accuracy} = \frac{TP + TN}{P + N}. \tag{5}$$

This metric alone could be biased and far from ideal, as a naïve predictor that predicts all examples as positive would already reach about $(14k + 0)/(14k + 6k) = 70\%$ classification accuracy. To complement such a metric when evaluating models on imbalanced testing sets, an alternative balanced accuracy metric [78] can be adopted:

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{TP}{P}\right) + \frac{1}{2}\left(\frac{TN}{N}\right). \tag{6}$$

Balanced accuracy equally considers the classification performance on different classes [78], [79]. While the overall accuracy in (5) offers an intuitive sense of correctness by reporting the proportion of correctly classified samples, the balanced accuracy in (6) combines the prevalence-independent statistics of sensitivity and specificity. A low balanced accuracy will be observed if a given classifier tends to predict only the dominant class. For the naïve predictor mentioned above, the balanced accuracy would give a proper number

Table 1. An overview of typical evaluation criteria.

Method	Formula	Remarks
Overall accuracy	$\frac{TP+TN}{P+N}$	Accounting for the proportion of correctly classified samples.
Balanced accuracy	$\frac{1}{2}\frac{TP}{P}+\frac{1}{2}\frac{TN}{N}$	Averaging precision and true negative prediction for imbalanced distribution.
PR curve	$p=\frac{TP}{TP+FP}, r=\frac{TP}{TP+FN}$	Measuring the relationship between precision and recall.
Euclidean distance	$\sqrt{\sum_i (Y_i-\hat{Y}_i)^2}$	Measuring the difference between the ground-truth score and aesthetic ratings. Y: ground-truth score, $\hat{Y}$: predicted score.
Correlation ranking	$\frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}$	Measuring the statistical dependence between the ranking of aesthetic prediction and ground truth. rg_X, rg_Y: rank variables, σ: standard deviation, cov: covariance.
ROC curve	$tpr=\frac{TP}{TP+FN}, fpr=\frac{FP}{FP+TN}$	Measuring model performance change by true positive rate and false positive rate when the binary discrimination threshold is varied.
Mean AP	$\frac{1}{n}\sum_i^n (\text{precision}(i)\times\Delta\text{recall}(i))$	The averaged AP values, based on precision and recall. precision(i) is calculated among the first i predictions, Δrecall(i): change in recall.

TP: true positive, *TN*: true negative, *P*: total positive, *N*: total negative, *FP*: false positive, *FN*: false negative, *tpr*: true positive rate, *fpr*: false positive rate.

Table 2. The methods evaluated on the CUHK-PQ data set.

Method	Data Set	Metric	Result	Training–Testing Remarks
Su et al. (2011) [72]	CUHK-PQ	Overall accuracy	92.06%	1,000 training, 3,000 testing
Marchesotti et al. (2011) [47]	CUHK-PQ	Overall accuracy	89.90%	50–50 split
Zhang et al. (2014) [67]	CUHK-PQ	Overall accuracy	90.31%	50–50 split, 12,000 subset
Dong et al. (2015) [50]	CUHK-PQ	Overall accuracy	91.93%	50–50 split
Tian et al. (2015) [54]	CUHK-PQ	Overall accuracy	91.94%	50–50 split
Zhang et al. (2016) [57]	CUHK-PQ	Overall accuracy	88.79%	50–50 split, 12,000 subset
Wang et al. (2016) [53]	CUHK-PQ	Overall accuracy	92.59%	4:1:1 partition
Lo et al. (2012) [66]	CUHK-PQ	Area under ROC curve	0.93	50–50 split
Tang et al. (2013) [45]	CUHK-PQ	Area under ROC curve	0.9209	50–50 split
Lv et al. (2016) [51]	CUHK-PQ	Mean AP	0.879	50–50 split

Table 3. The methods evaluated on the AVA data set.

Method	Data Set	Metric	Result	Training–Testing Remarks
Marchesotti et al. (2013) [48]	AVA	ROC curve	*tpr*: 0.7, *fpr*: 0.4	Standard partition
AVA handcrafted features (2012) [49]	AVA	Overall accuracy	68.00%	Standard partition
Spatial pyramid pooling (SPP) (2015) [24]	AVA	Overall accuracy	72.85%	Standard partition
RAPID (full method) (2014) [23]	AVA	Overall accuracy	74.46%	Standard partition
Peng et al. (2016) [52]	AVA	Overall accuracy	74.50%	Standard partition
Kao et al. (2016) [58]	AVA	Overall accuracy	74.51%	Standard partition
RAPID (improved version) (2015) [55]	AVA	Overall accuracy	75.42%	Standard partition
DMA-net (2015) [24]	AVA	Overall accuracy	75.41%	Standard partition
Kao et al. (2016) [59]	AVA	Overall accuracy	76.15%	Standard partition
Wang et al. (2016) [53]	AVA	Overall accuracy	76.94%	Standard partition
Kong et al. (2016) [25]	AVA	Overall accuracy	77.33%	Standard partition
BDN (2016) [56]	AVA	Overall accuracy	78.08%	Standard partition
Zhang et al. (2014) [67]	AVA	Overall accuracy	83.24%	10% subset, 12.5k*2
Dong et al. (2015) [50]	AVA	Overall accuracy	83.52%	10% subset, 19k*2
Tian et al. (2016) [54]	AVA	Overall accuracy	80.38%	10% subset, 20k*2
Wang et al. (2016) [53]	AVA	Overall accuracy	84.88%	10% subset, 25k*2
Lv et al. (2016) [51]	AVA	Mean AP	0.611	10% subset, 20k*2

indication of $0.5 \times (14k/14k) + 0.5 \times (0k/6k) = 50\%$ performance on AVA.

In this regard, in the following sections where we discuss our findings on a proposed strong baseline, we report both overall classification accuracy and balanced accuracy to get a more reasonable measure of baseline performance.

Experiments on deep-learning settings

It is evident from Table 3 that deep learning-based approaches dominate the performance of image aesthetic assessment. The effectiveness of learned deep features in this task has motivated us to take a step back to consider how a CNN works to understand the aesthetic quality of an image. It is worth noting that training a robust deep aesthetic scoring model is nontrivial, and often we found that the devil is in the details. To this end, we design a set of systematic experiments based on a baseline one-column CNN and a two-column CNN, and evaluate different settings from minibatch formation to complex multicolumn architecture. The results are reported on the widely used AVA data set.

We observe that by carefully training the CNN architecture, the two-column CNN baseline reaches comparable or

Table 4. The methods evaluated on other data sets.

Method	Data Set	Metric	Result
Tong et al. (2004) [20]	29,540-image private set	Overall accuracy	95.10%
Datta et al. (2006) [21]	3,581-image private set	Overall accuracy	75%
Sun et al. (2009) [38]	600-image private set	Euclidean distance	3.5135
Wong et al. (2009) [63]	3,161-image private set	Overall accuracy	79%
Bhattacharya (2010, 2011) [43], [64]	~650-image private set	Overall accuracy	86%
Li et al. (2010) [70]	500-image private set	Residual sum-of-squares error	2.38
Wu et al. (2010) [65]	10,800-image private set from Flickr	Overall accuracy	~83%
Dhar et al. (2011) [44]	16,000-image private set from DPChallenge	PR curve	–
Nishiyama et al. (2011) [41]	12,000-image private set from DPChallenge	Overall accuracy	77.60%
Lo et al. (2012) [42]	4,000-image private set	ROC curve	*tpr*: 0.6, *fpr*: 0.3
Yeh et al. (2012) [46]	309-image private set	Kendalls Tau-b measure	0.2812
Aydin et al. (2015) [62]	955-image subset from DPChallenge.com	Human survey	–
Yin et al. (2012) [73]	13,000-image private set from Flickr	Overall accuracy	81%
Lienhard et al. (2015) [71]	Human Face Scores 250-image data set	Overall accuracy	86.50%
Sun et al. (2015) [74]	1,000-image Chinese handwriting	Euclidean distance	–
Kong et al. (2016) [25]	AADB data set	Spearman ranking	0.6782
Zhang et al. (2016) [57]	PNE	Overall accuracy	86.22%

even better performance than state-of-the-art methods, and the one-column CNN baseline acquires the strong capability to suppress false-positive predictions while having competitive classification accuracy. We hope the experimental results will facilitate the design of future deep-learning models for image aesthetic assessment.

Formulation and the base CNN structure

The supervised CNN learning process involves a set of training data $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$ from which a nonlinear mapping function $f: X \rightarrow Y$ is learned through backpropagation [33]. Here, $\boldsymbol{x}_i$ is the input to the CNN and $y_i \in \mathbb{T}$ is its corresponding ground-truth label. For the task of binary classification, $y_i \in \{0,1\}$ is the aesthetic label corresponding to image $\boldsymbol{x}_i$. The convolutional operations in such a CNN can be expressed as

$$F_k(X) = \max\left(\mathbf{w}_k * F_{k-1}(X) + \mathbf{b}_k, 0\right), k \in \{1,2,...,D\}, \quad (7)$$

where $F_0(X) = X$ is the network input and D is the depth of the convolutional layers. The operator $*$ denotes the convolution operation. The operations in the D' fully connected layers can be formulated in a similar manner. To learn the $(D + D')$ network weights $\mathbf{W}$ using the standard backpropagation with stochastic gradient descent, we adopt the cross-entropy classification loss, which is formulated as

$$L(\mathbf{W}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t}\{t \log p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{W}) + (1-t)\log\left(1 - p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{W})\right) + \phi(\mathbf{W})\} \quad (8)$$

$$p(\hat{y}_i = t \mid \boldsymbol{x}_i; \mathbf{w}_t) = \frac{\exp\left(\mathbf{w}_i^T \boldsymbol{x}_i\right)}{\sum_{t' \in \mathbb{T}} \exp\left(\mathbf{w}_t^T \boldsymbol{x}_i\right)}, \quad (9)$$

where $t \in \mathbb{T} = \{0,1\}$ is the ground truth. This formulation is in accordance with prior successful model frameworks, such as AlexNet [75] and VGG-16 [80], which are also adopted as the base network in some of our reviewed approaches.

The original last fully connected layer of these two networks is for the 1,000-class ImageNet object recognition challenge. For aesthetic quality classification, a two-class aesthetic classification layer to produce a softmax predictor is needed [see Figure 9(a)]. Following typical CNN approaches, the input size is fixed to $224 \times 224 \times 3$, which is cropped from globally warped $256 \times 256 \times 3$ images. Standard data augmentation, such as mirroring, is performed. All of the baselines are implemented based on the Caffe package [81]. For clarity of presentation in the following sections, we name all of our fine-tuned baselines Deep Aesthetic Net (DAN), with the corresponding suffix.

Training from scratch versus fine-tuning

Fine-tuning from a trained CNN has been proven in [36] and [83] to be an effective initialization approach. The RAPID base network [23] uses global image patches and trains a network structure from scratch that is similar to AlexNet. For a fair comparison of similar-depth networks, we first select AlexNet pretrained with the ILSVRC-2012 training set (1.2 million images) and fine-tune it with the AVA training partition. As

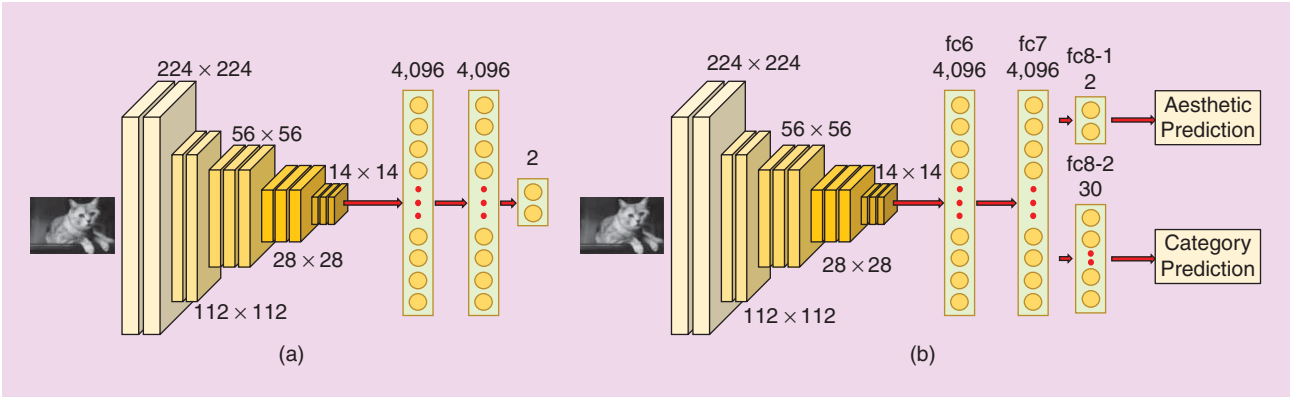


FIGURE 9. (a) The structure of the chosen base network for our systematic study on aesthetic quality classification. (b) The structure of the one-column CNN baseline with multitask learning [49].

Table 5. Training from scratch versus fine-tuning.

Method	Balanced Accuracy	Overall Accuracy
RAPID (global) [23]	–	67.8
DAN-1 (fine-tuned from AlexNet)	68.0	71.3
DAN-1 (fine-tuned from VGG-16)	72.8	74.1

Using a one-column CNN baseline (DAN-1) fine-tuned on AlexNet and VGG-16, both of which are pretrained on the ImageNet data set. The authors in [23] have not released detailed classification results.

Table 6. The effects of minibatch formation.

Minibatch Formation	Balanced Accuracy	Overall Accuracy
DAN-1 (randomly sampled)	70.39	77.65
DAN-1 (balanced formation)	72.82	74.06

Using a one-column CNN baseline (DAN-1) with VGG-16 as the base network.

shown in Table 5, fine-tuning from the vanilla AlexNet yields better performance than simply training the RAPID base network from scratch. Moreover, the DAN model fine-tuned from VGG-16 [see Figure 9(a)] yields the best performance in both balanced accuracy and overall accuracy. It is worth pointing out that other more recent and deeper models, such as ResNet [84], Inception-ResNet [85], and PolyNet [86], could serve as pretrained models. Nevertheless, owing to the typically small size of aesthetic data sets, precautions need be taken during the fine-tuning process. Plausible methods include freezing some earlier layers to prevent overfitting [83].

Minibatch formation

Minibatch formation directly affects the gradient direction toward which stochastic gradient descent brings down the training loss in the learning process. We consider two types of minibatch formation and reveal the impact of this difference on image aesthetic assessment.

Random sampling

By randomly selecting examples for minibatches [87], [88], we select from a distribution of the training partition. Since the number of positive examples in the AVA training partition is almost twice that of the negative examples [Figure 4(b)], models trained with such minibatches may bias toward predicting positives.

Balanced formation

Another approach is to enforce a balanced number of positives and negatives in each of the minibatches, i.e., for each iteration of backpropagation, the gradient is computed from a balanced number of positive examples and negative examples.

Table 6 compares the performance of these two strategies. We observe that although the model fine-tuned with randomly sampled minibatches reaches a higher overall accuracy, its performance is inferior to the one fine-tuned with balanced minibatches, as evaluated using balanced accuracy. To keep track of both true-positive prediction rates and true-negative prediction rates, balanced accuracy is adopted to measure the model robustness on the data imbalance issue. Network fine-tuning in the rest of the experiments is performed with balanced minibatches, unless otherwise specified.

Triplet pretraining and multitask learning

Apart from directly training using the given training data pairs $\{\boldsymbol{x}_i, y_i\}_{i\in[1,N]}$, one could utilize richer information inherent in the data or auxiliary sources to enhance the learning performance. We discuss two popular approaches next.

Pretraining using triplets

The triplet loss is inspired by Dimensionality Reduction by Learning an Invariant Mapping [89] and large margin nearest neighbor [90]. It is widely used in many recent vision studies [79], [91]–[93] and aims to bring data of the same class closer while moving data of different classes further away. This loss is particularly suitable to our task; i.e., the absolute aesthetic score of an image is arguably subjective, but the general relationship that beautiful images are close to each other while the opposite images should be apart is obvious.

Table 7. Triplets pretraining and multitask learning.

Methods	Balanced Accuracy	Overall Accuracy
DAN-1	72.82	74.06
DAN-1 (triplet pretrained)	73.29	75.32
DAN-1 (multitask–aesthetic and category)	73.39	75.36
DAN-1 (triplet pretrained + multitask)	**73.59**	74.42

Using a one-column CNN baseline (DAN-1) with VGG-16 as the base network. Balanced minibatch formation is used.

To enforce such a relationship in an aesthetic embedding, one needs to generate minibatches of triplets for deep feature learning, i.e., an anchor $\boldsymbol{x}$, a positive instance $\boldsymbol{x}_{+ve}$ of the same class, and a negative instance $\boldsymbol{x}_{-ve}$ of a different class. Furthermore, we found it useful to constrain each image triplet to be selected from the same image category. In addition, we observed better performance by introducing triplet loss in the pretraining stage and continuing with conventional supervised learning on the triplet-pretrained model. Table 7 shows that the DAN model pretrained with triplets gives better performance.

We further visualize some categories in the learned aesthetic embedding space in Figure 10. It is interesting to observe that the embedding learned with triplet loss demonstrates much better aesthetic grouping in comparison to that without the use of triplet loss.

Multitask learning with image category prediction

Can aesthetic prediction be facilitated provided that a model understand to which category the image belongs? Following the work in [94], where auxiliary information is used to regularize the learning of the main task, we investigate the potential benefits of using image categories as an auxiliary label in training the aesthetic quality classifier.

Specifically, given an image labeled with main task label y, where $y = 0$ for low-quality images and $y = 1$ for high-quality ones, we provide an auxiliary label $c \in C$ denoting one of the image categories, such as animals, landscape, portraits, and so forth. In total, we include 30 image categories. To learn a classifier for the auxiliary class, a new fully connected layer is attached to the fc7 of the vanilla VGG-16 structure to predict a softmax probability for each category class. The modified one-column CNN baseline architecture is shown in Figure 9(b). The loss function in (8) is now changed to

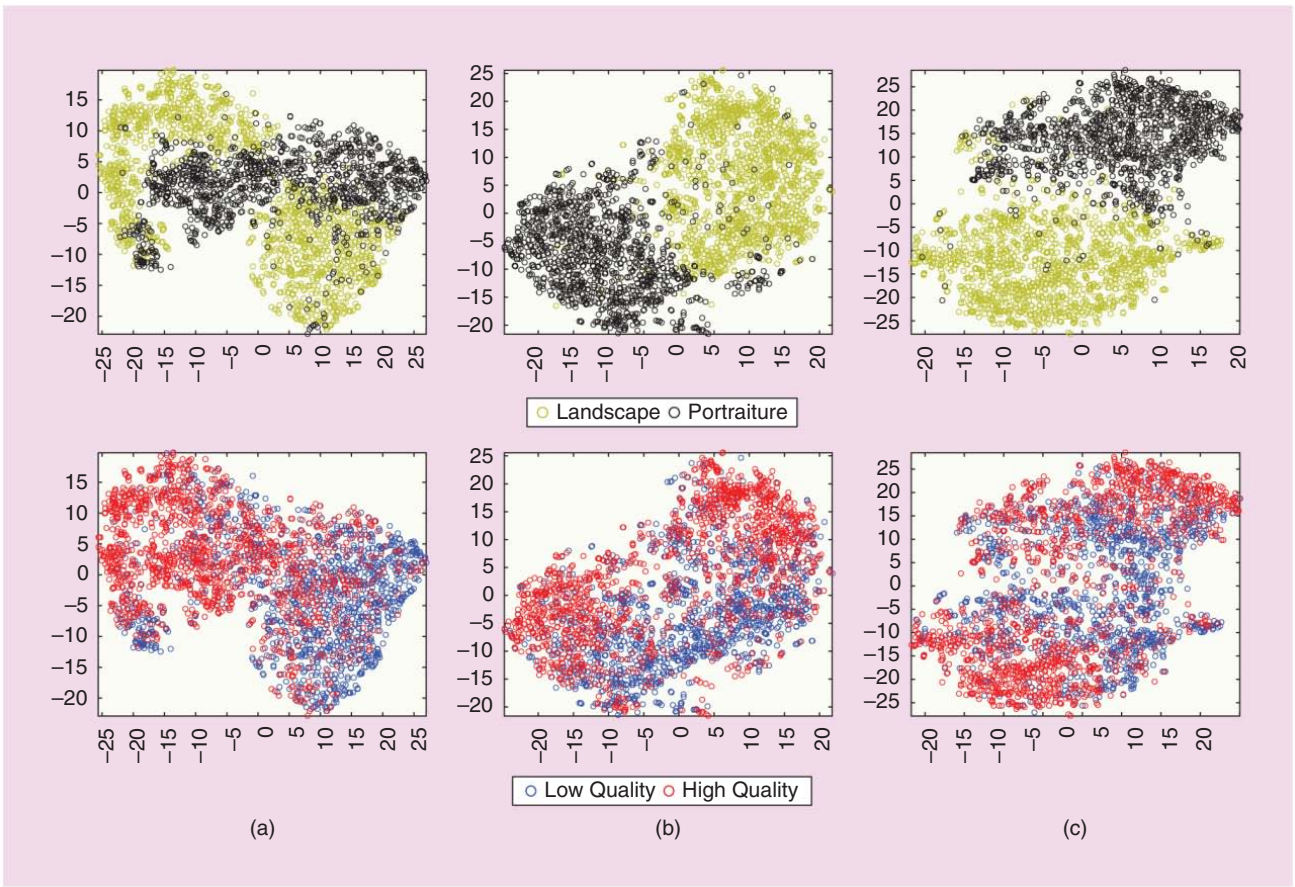


FIGURE 10. Aesthetic embeddings of AVA images (testing partition) learned by triplet loss, visualized using t-SNE [84]: (a) ordinary supervised learning without triplet pretraining and multitask learning, (b) triplet pretrained, and (c) combined triplet pretraining and multitask learning. t-SNE: t-distributed stochastic neighbor embedding.

$$L_{\text{multitask}} = L(\mathbf{W}) + L_{\text{aux}}(\mathbf{W}_c), \tag{10}$$

$$L_{\text{aux}}(\mathbf{W}_c) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\{t_c \log p(\hat{y}_c^{\text{aux}} = t_c \mid \boldsymbol{x}_i; \mathbf{W}_c) + (1 - t_c)\log p(\hat{y}_c^{\text{aux}} = t_c \mid \boldsymbol{x}_i; \mathbf{W}_c)) + \phi(\mathbf{W}_c)\}, \tag{11}$$

where $t_c \in \{0,1\}$ is the binary label corresponding to each auxiliary class $c \in C$ and $\hat{y}_c^{\text{aux}}$ is the auxiliary prediction from the network. Solving the above loss function, the DAN model performance from this multitask learning strategy is observed to have surpassed the previous one (Table 7). It is worth noting that the category annotation of the AVA-training partition is not complete, with about 25% of the images not having categories labeled. For those training instances without categories labeled, the auxiliary loss $L_{\text{aux}}(\mathbf{W}_c)$ due to missing labels is ignored.

The absolute aesthetic score of an image is arguably subjective, but the general relationship that beautiful images are close to each other while the opposite images should be apart is obvious.

Triplet pretraining + multitask learning

Combining triplet pretraining and multitask learning, the final one-column CNN baseline reaches a balanced accuracy of 73.59% on the challenging task of aesthetic classification. The results for different fine-tuning strategies is summarized in Table 7.

Discussion

Note that it is nontrivial to boost the overall accuracy at the same time as we try not to overfit the baseline to a certain data distribution. Still, compared with other released results in Table 8, with careful training, a one-column CNN baseline yields a strong capability of rejecting false positives while attaining a reasonable overall classification accuracy. We show some qualitative classification results as follows.

Figures 11 and 12 show the qualitative results of aesthetic classification by the one-column CNN baseline, using DAN-1 (triplet pretrained + multitask). Note that these examples are correctly classified neither by BDN [56] nor by DMA-net [24]. False-positive test examples (Figure 13) by the DAN-1 baseline still show a somewhat high-quality image trend, with high color contrast or depth of field, while false-negative testing examples (Figure 14) mostly reflect low image tones. Both quantitative and qualitative results suggest the importance of minibatch formation and fine-tuning strategies.

Multicolumn deep architecture

State-of-the-art approaches [23], [24], [55], [56] for image aesthetic classification typically adopt multicolumn CNNs (Figure 7) to enhance the learning capacity of the model. In particular, these approaches benefit from learning multiscale image information (e.g., global image versus local patches) or utilizing image semantic information (e.g., image styles). To incorporate insights from previous successful approaches, we prepared another two-column CNN baseline (DAN-2) (see Figure 15) with a focus on the more apparent approach of using local image patches as a parallel input column. Both [23] and [24] utilize CNNs trained with local image patches as alternative columns in their multibranch network, with performance evaluated using overall accuracy. For fair comparison, we prepared local image patches of size $224 \times 224 \times 3$ following [23] and [24], and we fine-tuned one DAN-1 model from the vanilla VGG-16 (ImageNet) with such local patches. Another branch is the original DAN-1 model, fine-tuned with globally warped input by triplet pretraining and multitask learning (see the section "Triplet Pretraining and Multitask Learning"). We performed separate experiments where minibatches of these local image patches were taken from either random sampling or the balanced formation.

As shown in Table 8, the DAN-1 model fine-tuned with local image patches performs less well under the metric of balanced accuracy compared to the original DAN-1 model fine-tuned with globally warped input in both random minibatch learning and balanced minibatch learning. We conjecture that local patches contain

Table 8. A comparison of aesthetic quality classification between our proposed baselines and previous state-of-the-art methods on the canonical AVA testing partition.

Previous Work	Balanced Accuracy	Overall Accuracy
AVA handcrafted features (2012) [49]	–	68.00
SPP (2015) [24]	–	72.85
RAPID (full method) (2014) [23]	–	74.46
Peng et al. (2016) [52]	–	74.50
Kao et al. (2016) [58]	–	74.51
RAPID (improved version) (2015) [55]	61.77	75.42
DMA-net (2015) [24]	62.80	75.41
Kao et al. (2016) [59]	–	76.15
Wang et al. (2016) [53]	–	76.94
Kong et al. (2016) [25]	–	77.33
Mai et al. (2016) [26]	–	77.40
BDN (2016) [56]	67.99	78.08
Proposed Baseline Using Random Minibatches		
DAN-1 (VGG-16, AVA global warped input)	70.39	77.65
DAN-1 (VGG-16, AVA local patches)	68.70	77.60
Two-column DAN-2	69.45	**78.72**
Proposed Baseline Using Balanced Minibatches		
DAN-1 (VGG-16, AVA global warped input)	**73.59**	74.42
DAN-1 (VGG-16, AVA local patches)	71.40	75.8
Two-column DAN-2	73.51	75.96

The authors of [23]–[26], [49], [52], [53], [55], [58], and [59] have not released detailed results.

FIGURE 11. Some positive examples (high-quality images) that are wrongly classified by BDN and DMA-net but correctly classified by the DAN-1 baseline [49].

FIGURE 12. Some negative examples (low-quality images) that are wrongly classified by BDN and DMA-net but correctly classified by the DAN-1 baseline [49].

FIGURE 13. Some examples with a negative ground truth that are wrongly classified by the DAN-1 baseline. High color contrast or depth of field is observed in these testing cases [49].

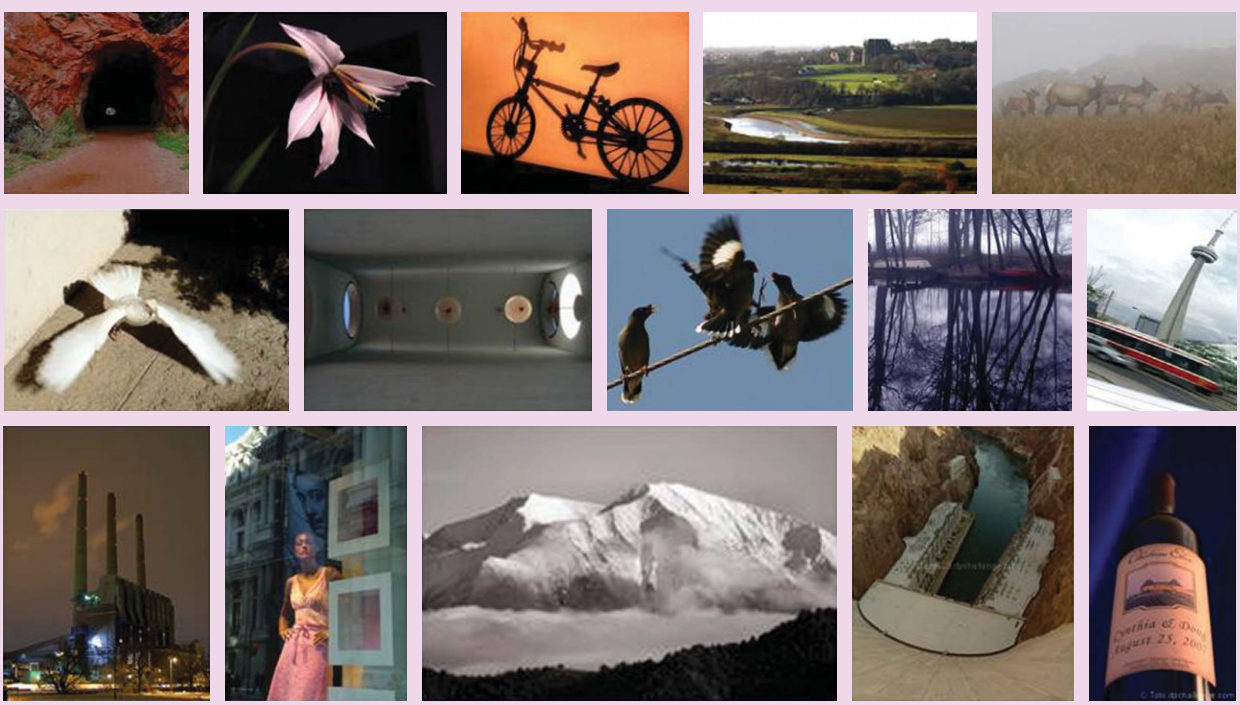


FIGURE 14. Some examples with a positive ground truth that are wrongly classified by the DAN-1 baseline. Most of these images are of low image tones [49].

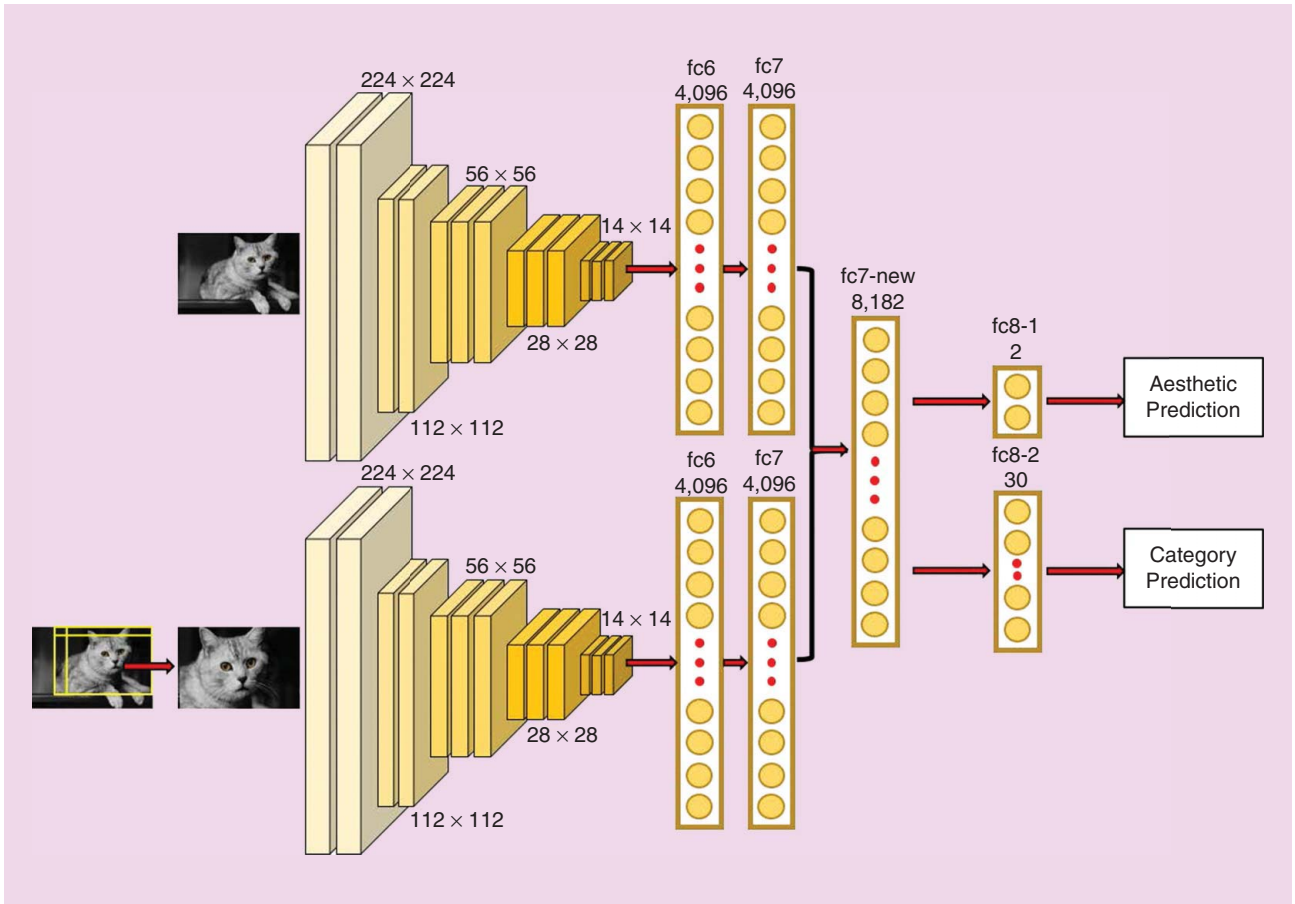


FIGURE 15. The structure of the two-column CNN baseline with multitask learning [49].

no global and compositional information as compared to globally warped input. Nevertheless, such a drop in accuracy is not observed under the overall accuracy metric.

We next evaluated the two-column CNN baseline DAN-2 using the DAN-1 model fine-tuned with local image patches and the one fine-tuned with globally warped input. We have two variants here, depending on whether we employ random or balanced minibatches. We observed that DAN-2 trained with random minibatches attains the highest overall accuracy on the AVA standard testing partition compared to the previous state-of-the-art methods (see Table 8). (Some other works [50], [54], [95]–[97] on AVA data sets use only a small subset of images for evaluation, which is not directly comparable to the canonical state of the art on the AVA standard partition; see Table 3).

Interestingly, we observed the balanced accuracy of the two variants of DAN-2 degrades when compared to the respective DAN-1 trained on globally warped input. This observation raises the question of whether local patches necessarily benefit the performance of image aesthetic assessment. We analyzed the cropped local patches more carefully and found that these patches were inherently ambiguous. Thus, the model trained with such inputs could easily become biased toward predicting local patch input to be of high quality, which also explains the performance differences in the two complementary evaluation metrics.

Can aesthetic prediction be facilitated provided that a model understand to which category the image belongs?

Model depth and layer-wise effectiveness

Determining the aesthetics of images from different categories takes varying photographic rules. We understand that it is not easy to determine some image genres' aesthetic quality in general. It would be interesting to perform a layer-by-layer analysis and track to what degree a deep model has learned image aesthetics in its hierarchical structure. We conducted this experiment using the one-column CNN baseline DAN-1 (triplet pretrained + multitask). We used layer features generated by this baseline model and trained an SVM classifier to perform aesthetic classification on the AVA testing images and then evaluated the performance of different layer features across different image categories.

Features extracted from the convolutional layers of the model were aggregated into a convolutional Fisher representation, as done in [98]. Specifically, to extract features from the dth convolutional layer, note that the output feature maps of

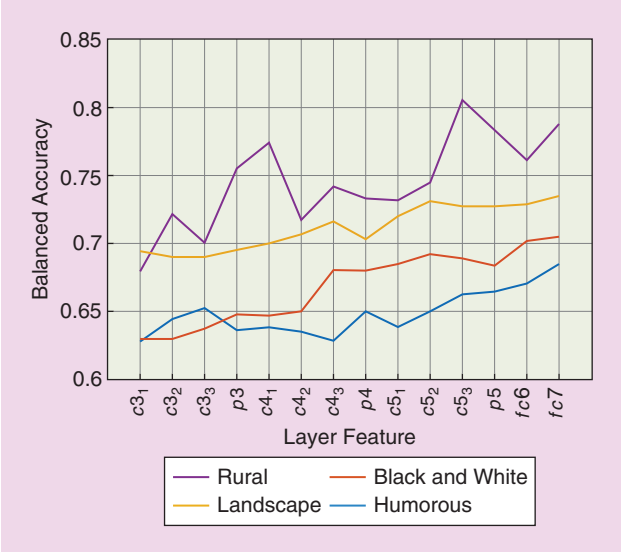


FIGURE 16. A layer-by-layer analysis showing the difficulties of understanding aesthetics across different categories. From the learned feature hierarchy and the classification results, we observe that image aesthetics in the landscape and rural categories can be judged reasonably by the proposed baselines, yet the more ambiguous humorous and black-and-white images are inherently difficult for the model to handle (see also Figure 17).

this dth layer are of size $w \times h \times K$, where $w \times h$ is the size of each of the K output maps. Denote M^k as the kth output map. Specifically, a point $M^k_{i,j}$ in output map M^k is computed from a local patch region L of the input image I using the forward propagation. By aligning all such points into a vector $\boldsymbol{v}_L = [M^1_{i,j}, M^2_{i,j}, \ldots, M^k_{i,j}, \ldots, M^K_{i,j}]$, we obtained the feature representation of the local patch region L. A dictionary codebook was created using GMM from all of the $\{\boldsymbol{v}_L\}_{L \in I_{\text{train}}}$, and an FV representation is subsequently computed using this codebook to describe an input image. The obtained convolutional Fisher representation is used for training SVM classifiers.

We compared features from layer conv3_1 to fc7 of the DAN-1 baseline and reported selected results that we find interesting in Figure 16. We obtained the following results:

1) *Model depth is important*: More abstract aesthetic representation can be learned in deeper layers. The performance of aesthetic assessment can generally be benefited from model depth. This observation aligns with that in general object recognition tasks.
2) *Different categories demand different model depths*: The aesthetic classification accuracy on images belonging to the black and white category are generally lower than the accuracy on images in the landscape category across all of the layer features. Sample classification results are shown in confusion matrix ordering (see Figure 17). High-quality black-and-white images show subtle details that should be considered when assessing their aesthetic level, whereas high-quality landscape images differentiate from those low-quality ones in a more apparent way. Similar observations are found, e.g., in the humorous and rural categories. The observation explains why it could be inherently difficult for the baseline model to judge whether images from some specific categories are aesthetically pleasing or not, revealing yet another challenge in the assessment of image aesthetics.

DAN-2 trained with random minibatches attains the highest overall accuracy on the AVA standard testing partition compared to the previous state-of-the-art methods.

From generic aesthetics to user-specific taste

Individual users may hold different opinions on the aesthetic quality of any single image. One may consider that all of the images in Figure 13 are of high quality to some extent, even though the average scores by the data set annotators say otherwise. Coping with individual aesthetic bias is a challenging problem. We may follow the idea behind transfer learning [83] and directly model the aesthetic preference of individual users by transferring the learned aesthetic features to fitting personal taste. In particular, we consider that the DAN-1 baseline network has already captured a sense of generic aesthetics in the aforementioned learning process; so to adapt to personal aesthetic preferences, one can include additional data sources for positive training samples that are user specific, such as the user's personal photographic album or the collection of photos that the user "liked" on social media. As such, our proposed baseline can be further fine-tuned with personal-taste data for individual users and become a personalized aesthetic classifier.

Image aesthetic manipulation

A task closely related to image aesthetic assessment is image aesthetics manipulation, the aim of which is to improve the aesthetic quality of an image. A full review of the techniques of image aesthetics manipulation in the literature is beyond the scope of this article. Still, we make an attempt to connect image aesthetic assessment to a broader topic surrounding image aesthetics by focusing on one of the major aesthetic enhancement operations, i.e., automatic image cropping.

Aesthetics-based image cropping

Image cropping improves the aesthetic composition of an image by removing undesired regions, increasing its aesthetic value. A majority of cropping schemes in the literature can be divided into three main approaches. Attention/saliency-based approaches [99]–[101] typically extract the primary subject region in the scene of interest according to attention scores or saliency maps as the image crops. Aesthetics-based approaches [102]–[104] assess the attractiveness of some proposed candidate crop windows with low-level image features and rules of photographic composition. However, simple handcrafted features are not robust for modeling the huge aesthetic space. The state-of-the-art method is the change-based approach proposed by Yan et al. [105], [106], which aims to

FIGURE 17. A layer-by-layer analysis of classification results using the best layer features on (a) black-and-white category images and (b) landscape category images [49].

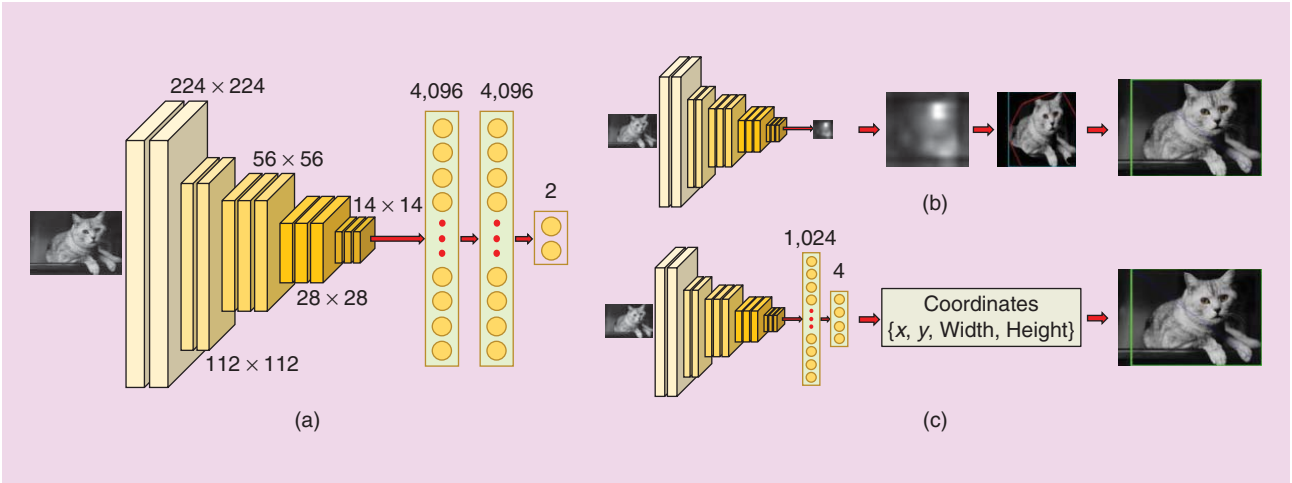


FIGURE 18. (a) The originally proposed one-column CNN baseline. (b) A tweaked CNN made by removing all of the fully connected layers. (c) A modified CNN incorporating a crop-regression layer to learn cropping coordinates [49].

account for what is removed and changed by cropping itself and trying to incorporate the influence of the starting composition of the initial image in the ending composition of the cropped image. This approach produces reasonable crop windows, but the time cost of producing an image crop is prohibitively expensive because of the time spent in evaluating large numbers of crop candidates.

Automatic thumbnail generation is also closely related to automatic image cropping. Huang et al. [107] target visual representativeness and foreground recognizability when cropping and resizing an image to generate its thumbnail. Chen et al. [108] aim at extracting the most visually important region as the image crop. Nevertheless, the aesthetics aspects of cropping are not taken into prime consideration in these approaches.

In the next section, we show that high-quality image crops can already be produced from the last convolutional layer of the aesthetic classification CNN. Optionally, this convolutional response can be utilized as the input to a cropping regression layer for learning more precise cropping windows from additional crop data.

High-quality image crops can already be produced from the last convolutional layer of the aesthetic classification CNN.

Plausible formulations based on deep models

Fine-tuning a CNN model for the task of aesthetic quality classification (see the "Experiments on Deep-Learning Settings" section) can be considered as a learning process in which the fine-tuned model tries to understand the metric of image aesthetics. We hypothesize that the same metric is applicable to the task of automatic image cropping. We discuss two possible variants as follows.

DAN-1 (original) without cropping data

Without utilizing additional image cropping data, a CNN such as the one-column CNN baseline DAN-1 can be tweaked to produce image crops with minor modifications, removing the fully connected layers. That leaves us with a neural network that is fully convolutional where the input can be of arbitrary size, as shown in Figure 18(b). The output of the last convolutional layer of the modified model is $14 \times 14 \times 512$ dimensional, where the 512 feature maps contain the responses/activations corresponding to the input. To generate the final image crop, we take an average of the 512 feature maps and resize it to the input image size. After that, a binary mask is generated by suppressing the feature map values below a threshold. The output crop window is produced by taking a rectangle convex hull from the largest connected region of this binary mask.

DAN-1 (regression) with cropping data

Alternatively, to include additional image cropping data $\{\boldsymbol{x}_i^{\text{crop}}, Y_i^{\text{crop}}\}_{i\in[1,N]}$, where $Y_i^{\text{crop}} = [x, y, \text{width}, \text{height}]$, we follow insights in [111] and add a window regression layer to learn a mapping from the convolutional response [see Figure 18(c)]. As such, we can predict a more precise cropping window by learning this extended regressor from such crop data by a Euclidean loss function:

$$L(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \| \hat{Y}_i^{\text{crop}} - Y_i^{\text{crop}} \|^2, \tag{12}$$

where $\hat{Y}_i^{\text{crop}}$ is the predicted crop window for input image $\boldsymbol{x}_i^{\text{crop}}$.

To learn the regression parameters for this additional layer, the image cropping data set by Yan et al. [105] is used for further fine-tuning. Images in the data set are labeled with ground-truth crops by professional photographers. Following the evaluation criteria in [105], a fivefold cross-validation approach is adopted for evaluating the model performance on

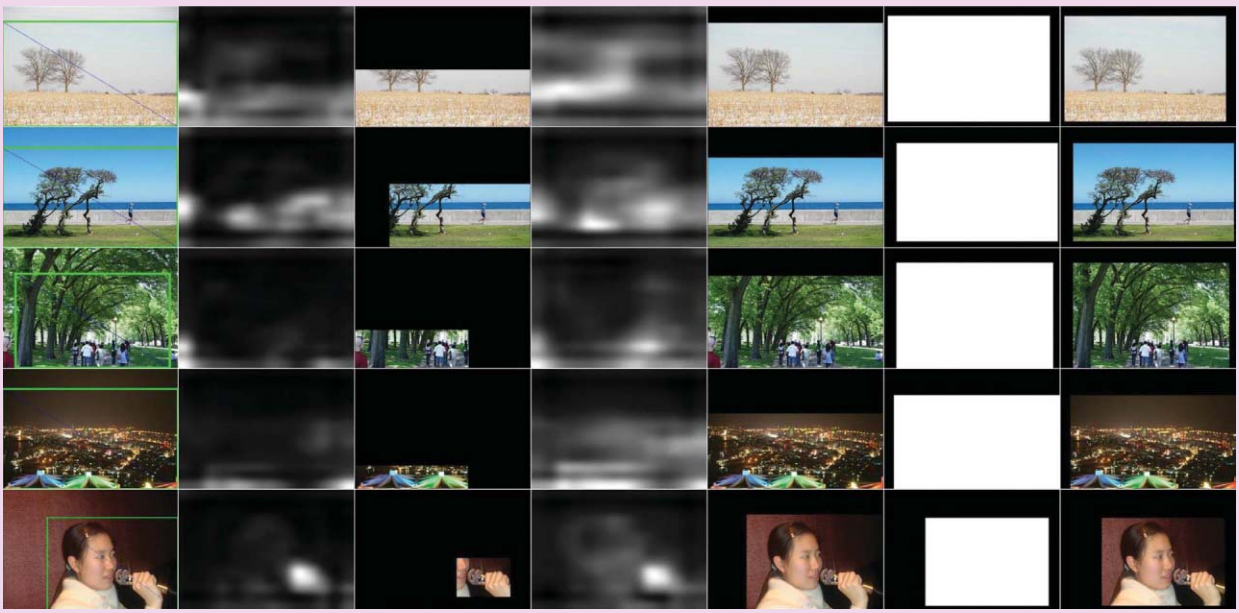


FIGURE 19. The layer response differences of the last convolutional layer. The images in each row correspond to (a) the input image with ground-truth crop, (b) the feature response of the vanilla VGG, (c) the image crops obtained via the feature responses of the vanilla VGG, (d) the feature response of the DAN-1 (original) model, (e) the image crops obtained via the DAN-1 (original) model, (f) the four-coordinates window estimated by the DAN-1 (regression) network, and (g) the cropped image generated by the DAN-1 (regression) [107].

all images in the data set. Note that there are only a few hundred images in each training fold; hence, a direct fine-tuning by simply warping the few hundred images of input to $224 \times 224 \times 3$ could be vulnerable to overfitting. To this end, we fix the weights in the convolutional layers of the DAN-1 (regression) network and learn only the weights for the crop window regression layers. Also, a systematic augmentation approach is adopted as follows. First, the input images are randomly jittered for a few pixels ($\times 5$), and mirroring is performed ($\times 2$). Second, we warp the images to have their longer side equal to 224, hence keeping their aspect ratio. We further downscale the images using a scale of $C \in \{50\%, 60\%, 80\%, 90\%\}$ ($\times 4$). The downscaled images are then padded back to 224×224 from $\{\text{top} - \text{left}, \text{top} - \text{right}, \text{bottom} - \text{left}, \text{bottom} - \text{right}\}$ ($\times 4$). Finally, we have direct input warping regardless of the aspect ratio ($\times 1$). In this manner, one training instance is augmented to $5 \times 2 \times (4 \times 4 + 1) = 170$ input instances. We fine-tune this modified CNN baseline with a learning rate of $10e^{-3}$, and the fine-tuning process converges at around the second epoch.

Aesthetics-based image cropping

As shown in Figure 19, we observe that the convolutional response of the vanilla VGG-16 (ImageNet) for object recognition typically finds a precise focus of the salient object in view, while the one-column CNN baseline, i.e., the DAN-1 (original) for aesthetic quality classification, outputs an aesthetically oriented salient region where both the object in view and its object composition are revealed. Compared to the cropping performance using the vanilla VGG-16, image crops from our DAN-1 (original) baseline already have the capability of removing unwanted regions while preserving the aesthetically salient part in view (see Figure 19). The modified CNN, i.e., the DAN-1 (regression), further incorporates aesthetic composition information in its crop window regression layer, which serves to refine the crop coordinates for more precise crop generation.

Following the evaluation settings in [105] and [106], we use the average overlap ratio and average boundary displacement error to quantify the performance of automatic image cropping. A higher overlap and a lower displacement between the generated crop and the corresponding ground truth indicate a more precise crop predictor. As shown in Table 9, directly using the DAN-1 (original) baseline responses to construct image crops already gains competitive cropping performance, while fine-tuning the DAN-1 (regression) with cropping data further boosts the performance and even surpasses the previous state-of-the-art method [105] on this data set, especially in terms of the boundary displacement error. Last but not least, it is worth noting that the CNN-based cropping approach takes merely ~0.2 s for generating an output image crop on a graphics processing unit and ~2 s on a central processing unit (compared to ~11 s on CPU in [105]).

Table 9. The performance on automatic image cropping.

Previous Work	*Photographer 1	Photographer 2	Photographer 3
Park et al. [111]	0.6034 (0.1062)	0.5823 (0.1128)	0.6085 (0.1102)
Yan et al. [108]	0.7487 (0.0667)	0.7288 (0.0720)	0.7322 (0.0719)
Wang et al. [112]	0.7823 (0.0623)	0.7697 (0.0617)	0.7725 (0.0701)
Yan et al. [107]	0.7974 (0.0528)	**0.7857** (0.0567)	0.7723 (0.0594)
Proposed Baselines			
Vanilla VGG-16 (ImageNet)	0.6971 (0.0580)	0.6841 (0.0618)	0.6715 (0.0613)
DAN-1 (original) (AVA training partition)	0.7637 (0.0437)	0.7437 (0.0493)	0.7360 (0.0495)
DAN-1 (regression) (cropping data fine-tuned)	**0.8059 (0.0310)**	0.7750 **(0.0375)**	**0.7725 (0.0377)**

*There are separate ground-truth annotations by three different photographers in the cropping data set of [107].
The first number is the average overlap ratio (higher is better). The second number (shown in parentheses) is the average boundary displacement error (lower is better).
Bold values signify the best performance by the corresponding methods.

Conclusion and potential directions

Models with competitive performance on image aesthetic assessment have been seen in the literature, yet the state of research in this field is far from saturated. Challenging issues include the ground-truth ambiguity due to neutral image aesthetics and how to effectively learn category-specific image aesthetics from the limited amount of auxiliary data information. Image aesthetic assessment can also benefit from an even larger volume of data, with richer annotations, where every single image is labeled by more users with diverse backgrounds. A large and more diverse data set will facilitate the learning of future models and potentially allow more meaningful statistics to be captured.

In this work, we systematically review major attempts on image aesthetic assessment in the literature and further propose an alternative baseline to investigate the challenging problem of understanding image aesthetics. We also discuss an extension of image aesthetic assessment to the application of automatic image cropping by adapting the learned aesthetic-classification CNN for the task of aesthetics-based image cropping. We hope that this survey can serve as a comprehensive reference source and inspire future research in understanding image aesthetics and fostering many potential applications.

Authors

Yubin Deng (dy015@ie.cuhk.edu.hk) received his B.Eng. degree (first-class honors) in information engineering from the Chinese University of Hong Kong in 2015. He is currently working toward his Ph.D. degree in the Department of Information Engineering, Chinese University of Hong Kong, with a Hong Kong Ph.D. Fellowship. His research interests include computer vision, pattern recognition, and machine learning. He was a Hong Kong Jockey Club Scholar in 2013–2014. He received the Professor Charles K. Kao Student Creativity Awards champion award in 2015.

Chen Change Loy (ccloy@ie.cuhk.edu.hk) received his B.Eng degree (first-class honors) from the University of Science, Malaysia, in 2005 and his Ph.D. degree in computer science from Queen Mary University of London, United Kingdom, in 2010. He is currently a research assistant professor in the Department of Information Engineering, Chinese University of Hong Kong. Previously, he was a postdoctoral researcher at Queen Mary University of London and Vision Semantics Ltd. His research interests include computer vision and pattern recognition, with a focus on facial analysis, deep learning, and visual surveillance. He serves as an associate editor of *IET Computer Vision Journal* and is a guest editor of *Computer Vision and Image Understanding*. He is a Member of the IEEE.

Xiaoou Tang (xtang@ie.cuhk.edu.hk) received his B.S. degree from the University of Science and Technology of China, Hefei, in 1990, his M.S. degree from the University of Rochester, New York, in 1991, and his Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1996. He is a professor in and the chair of the Department of Information Engineering, Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at Microsoft Research Asia from 2005 to 2008. His research interests include computer vision, pattern recognition, and video processing. He received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition 2009. He was a program chair of the IEEE International Conference on Computer Vision 2009, and he is an editor-in-chief of *International Journal of Computer Vision* and an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is a Fellow of the IEEE.

References

[1] M. Freeman, *The Complete Guide to Light and Lighting in Digital Photography* (A Lark Photography Book). New York: Sterling Publishing Company, 2007.

[2] J. Itten, *Design and Form: The Basic Course at the Bauhaus and Later*. New York: Wiley, 1975.

[3] B. London and J. Upton, *Photography*. London: Pearson, 2005.

[4] A. Chatterjee and O. Vartanian, "Neuroscience of aesthetics," *Ann. New York Acad. Sci.*, vol. 1369, no. 1, pp. 172–194, 2016.

[5] G. T. Fechner, *Vorschule der Aesthetik*, vol. 1. Wiesbaden, Germany: Breitkopf & Härtel, 1876.

[6] S. Zeki, "Clive Bell's 'significant form' and the neurobiology of aesthetics," *Frontiers Human Neurosci.*, vol. 7, p. 730, Nov. 2013.

[7] T. Ishizu and S. Zeki, "The brain's specialized systems for aesthetic and perceptual judgment," *Euro. J. Neurosci.*, vol. 37, no. 9, pp. 1413–1420, 2013.

[8] S. Brown, X. Gao, L. Tisdelle, S. B. Eickhoff, and M. Liotti, "Naturalizing aesthetics: Brain areas for aesthetic appraisal across sensory modalities," *Neuroimage*, vol. 58, no. 1, pp. 250–258, 2011.

[9] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The experience of emotion," *Annu. Rev. Psychol.*, vol. 58, pp. 373–403, Jan. 2007.

[10] L. F. Barrett and T. D. Wager, "The structure of emotion: Evidence from neuroimaging studies," *Current Directions Psychol. Sci.*, vol. 15, no. 2, pp. 79–83, 2006.

[11] S. Kühn and J. Gallinat, "The neural correlates of subjective pleasantness," *Neuroimage*, vol. 61, no. 1, pp. 289–294, 2012.

[12] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *Brit. J. Psychol.*, vol. 95, no. 4, pp. 489–508, 2004.

[13] A. Chatterjee, "Prospects for a cognitive neuroscience of visual aesthetics," *Bulletin Psychol. and the Arts*, vol. 4, no. 2, pp. 56–60, 2004.

[14] M. W. Greenlee and U. T. Peter, "Functional neuroanatomy of the human visual system: A review of functional MRI studies," in *Pediatric Ophthalmology, Neuro-Ophthalmology, Genetics*. Berlin: Springer, 2008, pp. 119–138.

[15] B. Wandell, S. Dumoulin, and A. Brewer, "Visual cortex in humans," *Encyclopedia of Neuroscience*, vol. 10, pp. 251–257, 2009.

[16] S. Zeki and J. Nash, *Inner Vision: An Exploration of Art and the Brain*. London: Oxford Univ. Press, 1999.

[17] P. Cavanagh, "The artist as neuroscientist," *Nature*, vol. 434, no. 7031, pp. 301–307, 2005.

[18] T. Ang, *Digital Photographer's Handbook*. London: Dorling Kindersley Publishing, 2002.

[19] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Boca Raton, FL: CRC, 2007.

[20] H. Tong, M. Li, H.-J. Zhang, J. He, and, and C. Zhang, "Classification of digital photos taken by photographers or home users," in *Advances in Multimedia Information Processing*, K. Aizawa, Y. Nakamura, and S. Satoh, Eds., *Lecture Notes in Computer Science*, vol. 3331. Berlin: Springer, 2004, pp. 198–205.

[21] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. European Conf. Computer Vision (ECCV)*. Berlin: Springer, 2006, pp. 288–301.

[22] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Computer Graphics Forum*, vol. 29, no. 2, pp. 469–478, 2010.

[23] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 457–466.

[24] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 990–998.

[25] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in Proc. *European Conf. Computer Vision (ECCV)*, 2016, pp. 662–679.

[26] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 497–506.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, 2011.

[29] A. Ebrahimi Moghadam, P. Mohammadi, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *Majlesi J. Elect. Eng.*, vol. 9, Mar. 2015.

[30] A. G. George and K. Prabavathy, "A survey on different approaches used in image quality assessment," *Int. J. Computer Sci. and Network Security*, vol. 14, no. 2, p. 78, 2014.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[32] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Processing and Quality Metrics for Consumer Electronics*, 2005, pp. 23–25.

[33] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, D. S. Touretzsky, Ed. Neural Information Processing Systems Foundation, 1989.

[34] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.

[35] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," in ACM *Trans. Graphics*, vol. 27, no. 3, article no. 73, 2008.

[36] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 576–584.

[37] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 419–426.

[38] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 541–544.

[39] J. You, A. Perkis, M. M. Hannuksela, and, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2009, pp. 561–564.

[40] Y. Luo, and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. European Conf, Computer Vision (ECCV)*, 2008, pp. 386–399.

[41] M. Nishiyama, T. Okabe, I. Sato, and, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 33–40.

[42] L.-Y. Lo and J.-C. Chen, "A statistic approach for photo quality assessment," in *Proc. IEEE Int. Conf. Information Security and Intelligence Control*, 2012, pp. 107–110.

[43] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 271–280.

[44] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1657–1664.

[45] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[46] M.-C. Yeh and Y.-C. Cheng, "Relative features for photo quality assessment," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2012, pp. 2861–2864.

[47] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2011, pp. 1784–1791.

[48] L. Marchesotti, F. Perronnin, and F. Meylan, "Learning beautiful (and ugly) attributes," in *Proc. British Machine Vision Conf. (BMVC)*, vol. 7, 2013, pp. 1–11.

[49] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2408–2415.

[50] Z. Dong, X. Shen, H. Li, and X. Tian, "Photo quality assessment with DCNN that understands image well," in *Proc. Int. Conf. Multimedia Modeling*, 2015, pp. 524–535.

[51] H. Lv and X. Tian, "Learning relative aesthetic quality with a pairwise approach," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 493–504.

[52] K.-C. Peng and T. Chen, "Toward correlating and solving abstract tasks using convolutional neural networks," in *Proc. IEEE Winter Conf. Applications Computer Vision (WACV)*, 2016, pp. 1–9.

[53] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Signal Process.: Image Commun.*, vol. 47, pp. 511–518, Sept. 2016.

[54] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[55] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[56] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," arXiv preprint arXiv:1601.04155, 2016.

[57] L. Zhang, "Describing human aesthetic perception by deeply-learned attributes from Flickr," arXiv preprint arXiv:1605.07699, 2016.

[58] Y. Kao, K. Huang, and S. Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Signal Process.: Image Commun.*, vol. 47, pp. 500–510, Sept. 2016.

[59] Y. Kao, R. He, and K. Huang, "Visual aesthetic quality assessment with multi-task deep learning," arXiv preprint arXiv:1604.04970, 2016.

[60] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *Proc. IEEE Int. Conf. Image Processing*, 2008, pp. 105–108.

[61] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* 2011, pp. 2206–2213.

[62] T. O. Ayd in, A. Smolic, and M. Gross, "Automated aesthetic analysis of photographic images," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 1, pp. 31–42, 2015.

[63] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2009, pp. 997–1000.

[64] S. Bhattacharya, R. Sukthankar, and M. Shah, "A holistic approach to aesthetic enhancement of photographs," *ACM Trans. Multimedia Computing, Commun., and Applicat.*, vol. 7S, no. 1, 2011.

[65] Y. Wu, C. Bauckhage, and C. Thurau, "The good, the bad, and the ugly: Predicting aesthetic image labels," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 1586–1589.

[66] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Assessment of photo aesthetics with efficiency," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2012, pp. 2186–2189.

[67] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1419–1429, 2014.

[68] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. European Conf. Computer Vision (ECCV) Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, 2004, pp. 1–2.

[69] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2007, pp. 1–8.

[70] C. Li, A. Gallagher, A. C. Loui, and T. Chen, "Aesthetic quality assessment of consumer photos with faces," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2010, pp. 3221–3224.

[71] A. Lienhard, P. Ladret, and A. Caplier, "Low level features for quality assessment of facial images," in *Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP)*, 2015, pp. 545–552.

[72] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 1213–1216.

[73] W. Yin, T. Mei, and C. W. Chen, "Assessing photo quality with geo-context and crowdsourced photos," in *Proc. IEEE Visual Communications and Image Processing Conf.*, 2012, pp. 1–6.

[74] R. Sun, Z. Lian, Y. Tang, and J. Xiao, "Aesthetic visual quality evaluation of Chinese handwritings," in *Proc. Int. Conf. Artificial Intelligence*, 2015, pp. 2510–2516.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges. L. Bottou, and K. Q. Weinberger, Eds. Neural Information Processing Systems Foundation, 2012, pp. 1097–1105.

[76] T. Joachims, "Training linear SVMs in linear time," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2006, pp. 217–226.

[77] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Int. Conf. Artificial Intelligence and Statistics*, Feb 2015, pp. 562–570.

[78] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. IEEE Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3121–3124.

[79] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 5375–5384.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[81] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," arXiv preprint arXiv:1408.5093, 2014.

[82] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Machine Learning Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[83] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27,* Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Neural Information Processing Systems Foundation, 2014, pp. 3320–3328.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 770–778.

[85] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261, 2016.

[86] X. Zhang, Z. Li, C. C. Loy, and D. Lin, "Polynet: A pursuit of structural diversity in very deep networks," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, in press.

[87] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2014, pp. 661–670.

[88] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.

[89] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.

[90] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Neural Information Processing Systems Foundation, 2005, pp. 1473–1480.

[91] B. Seguin, C. Striolo, F. Kaplan, et al. "Visual link retrieval in a database of paintings," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 753–767.

[92] Y. Wang and W. Deng, "Self-restraint object recognition by model based CNN learning," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2016, pp. 654–658.

[93] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1335–1344.

[94] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, 2016.

[95] Y. Kao, C. Wang, and K. Huang, "Visual aesthetic quality assessment with a regression model," in *Proc. Int. Conf. Image Processing*, 2015, pp. 1583–1587.

[96] E. Mavridaki and V. Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *Proc. Int. Conf. Image Processing*, 2015, pp. 887–891.

[97] Z. Dong and X. Tian, "Multi-level photo quality assessment with multi-view features," *Neurocomputing*, vol. 168, pp. 308–319, Nov. 2015.

[98] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, 2015.

[99] N. Jaiswal and Y. K. Meghrajani, "Saliency based automatic image cropping using support vector machine classifier," in *Proc. Int. Conf. Innovations Information, Embedded and Communication Systems*, 2015, pp. 1–5.

[100] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *Int. J. Comput. Vision*, vol. 104, no. 2, pp. 135–153, 2013.

[101] E. Ardizzone, A. Bruno, and G. Mazzola, "Saliency based image cropping," in *Proc. Int. Conf. Image Analysis and Processing*, 2013, pp. 773–782.

[102] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 669–672.

[103] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 291–300.

[104] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, 2013.

[105] J. Yan, S. Lin, S. B. Kang, and X. Tang, "Change-based image cropping with exclusion and compositional features," *Int. J. Comput. Vision*, vol. 114, no. 1, pp. 74–87, 2015.

[106] J. Yan, S. Lin, S. Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* 2013, pp. 971–978.

[107] J. Huang, H. Chen, B. Wang, and S. Lin, "Automatic thumbnail generation based on visual representativeness and foreground recognizability," in *Proc. IEEE Int. Conf. Computer Vision (ICCV),* 2015, pp. 253–261.

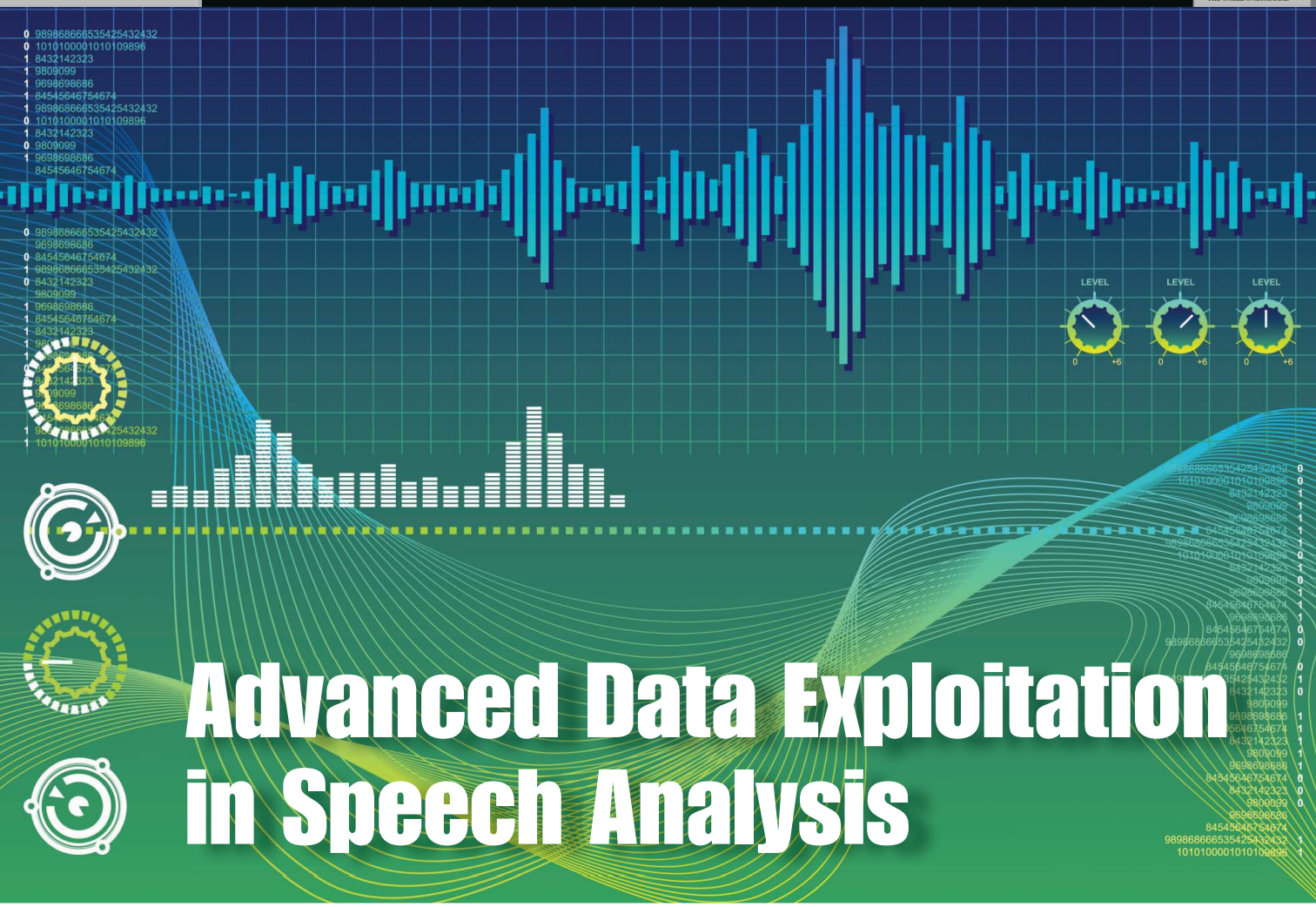
[108] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 507–515.

[109] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Modeling photo composition and its application to photo re-arrangement," in *Proc. Int. Conf. Image Processing*, 2012, pp. 2741–2744.

[110] P. Wang, Z. Lin, and R. Mech, "Learning an aesthetic photo cropping cascade," in *Proc. IEEE Winter Conf. Applications Computer Vision (WACV),* 2015, pp. 448–455.

[111] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.

SP



Advanced Data Exploitation in Speech Analysis

©ISTOCKPHOTO.COM/VIKTORUS

An overview

Zixing Zhang, Nicholas Cummins, and Björn Schuller

With recent advances in machine-learning techniques for automatic speech analysis (ASA)—the computerized extraction of information from speech signals—there is a greater need for high-quality, diverse, and very large amounts of data. Such data could be game-changing in terms of ASA system accuracy and robustness, enabling the extraction of feature representations or the learning of model parameters immune to confounding factors, such as acoustic variations, unrelated to the task at hand. However, many current ASA data sets do not meet the desired properties. Instead, they are often recorded under less than ideal conditions, with the corresponding labels sparse or unreliable.

In addressing these issues, this article provides a comprehensive overview of state-of-the-art ASA data exploitation techniques that have been developed to take advantage of knowledge gained from related but unlabeled or different data sources to improve the performance of a particular ASA task of interest. We first identify three primary data challenges: sparse, unreliable, and unmatched data. We then review the corresponding approaches. The conditions, advantages, and drawbacks of using a range of differing data-mining techniques are also discussed. Finally, other data challenges and potential future research directions in this field are presented.

Digital Object Identifier 10.1109/MSP.2017.2699358
Date of publication: 11 July 2017

Introduction to automatic speech analysis

ASA has long been regarded as one of the most vital areas in achieving natural and friendly human–machine interactions [1], [2]. The goal of ASA is to empower machines to automatically discern information of interest from human speech, e.g., identifying what is being said (the linguistic content), who is saying it (the speaker's identity), and how they are saying it (the paralinguistic content). More formally, typical ASA tasks in the literature include, but are not limited to,

- automatic speech recognition (ASR), which aims to extract linguistic content (e.g., words) by recognizing and translating spoken speech

- speaker identification/verification, which targets obtaining the speaker's identity from speech signals
- computational paralinguistics, which attempts to distill nonlinguistic information mainly concerning the speaker's short-term states (e.g., emotions), medium-term states (e.g., health condition and attitude), and long-term traits (e.g., personality, age, and gender) from spoken speech.

A serious obstacle to the broad application of ASA is the lack of sufficiently labeled data in terms of both quantity and quality. For example, many available computational paralinguistics corpora contain only a few hours of audio data at most [3]. Similarly for ASR, many of the world's languages are in a low-resource setting, where the electronic speech resources and linguistic expertise are lacking. According to a 2010 United Nations Educational, Scientific, and Cultural Organization report [4], approximately 2,500 languages are in danger of becoming extinct. In this scenario, it is exceptionally difficult to obtain a large-scale amount of transcribed speech data to perform reliable ASR.

A serious obstacle to the broad application of ASA is the lack of sufficiently labeled data in terms of both quantity and quality.

The requirement for large-scale labeled data is not new in machine leaning. Prevailing paradigms are often conducted in a supervised manner, and a substantial increase in the amount of available training data usually brings encouraging performance improvements [5]. Because of the advancement of deep-learning technologies [6], [7], this need for data has become more compelling than ever. Deep-learning models are often designed with millions of parameters, and, if trained with insufficient amounts of data, are vulnerable to being trapped in a locally optimized minimum, resulting in overfitting to the training data [6]. When sufficiently trained, however, deep models reach unprecedented levels of performance. For example, Amodei et al. [7] utilized approximately 12,000 and 9,000 h of speech data to model English and Mandarin ASR systems, respectively, by employing deep-learning models with more than 35 million trainable parameters, achieving a performance breakthrough that exceeds the capability of even human perception. Sufficient and reliably labeled data, when available, provide the opportunity to train robust ASA models whose resulting recognition is largely invariant in the face of the abundance of acoustic variations naturally present in speech data.

Opportunities

Traditionally, tasks such as data collection and annotation have been performed by small groups of experts in a laboratory setting. This conventional work paradigm is often tedious, time consuming, and costly. However, the ongoing information and communication technologies revolution and related technologies, such as the Internet of Things (IoT) and cloud computing, are providing us with opportunities to exploit larger amounts of speech data in more effective ways than ever before.

The IoT, as a global infrastructure of the information society, is expected to offer advanced services (i.e., data collection) by interconnecting a wide variety of contemporary recording devices, such as smartphones, wearable devices, and tablets. Furthermore, as these devices often have microphones, social media apps, and Internet connectivity, they can be considered distributed sensors or entryways for speech collection and processing. Thus, the advance of Internet technologies and the ubiquity of smart devices can drastically reduce the cost and time associated with collecting and processing speech data.

Cloud computing, or Internet-based computing, is expected to provide an on-demand computing resource. Thus, it gives an opportunity to store, access, and analyze the volume of speech data generated by the distributed devices mentioned previously. Cloud computing has been shown not only to minimize the costs associated with an ever-increasing demand for greater computational resources but also to reduce the cost associated with infrastructure maintenance and user access. Motivated by these advantages, most major speech technology providers have already shifted their primary research and application attention from embedded systems to cloud computing platforms.

Generalized automatic speech analysis: Problem statement and notation

The aforementioned technologies provide great potential to generate and process a large amount of speech data. However, there are three main challenges—data sparsity, unreliability, and nonmatching (Figure 1)—that limit the dissemination of these data in research and industry. Before formally defining these challenges, we first overview the generalized mathematical problem statement and notation commonly used in both ASA and throughout the remainder of this article.

First, let us define a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ that comprises a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where X denotes a set of feature vectors, i.e., $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}$; while $P(X)$ indicates the distribution of X in $\mathcal{X}$. In the case that each feature vector $\mathbf{x}$ consists of d attributes, i.e., $\mathbf{x} = \{x_1, \ldots, x_d\}$, $\mathcal{X}$ is a d-dimensional space. The most commonly used feature space $\mathcal{X}$ for ASA is arguably the Mel-frequency cepstral coefficients (MFCCs) that are extracted via filtering a speech frame by a bank of nonlinear bandpass filters (Mel filters) whose frequency response is based on the cochlea of the human auditory system [8]. Other exemplary feature spaces include the *i*-vector representation often used for speaker identification/verification [9], and mixed brute force feature representations, such as the broadly used ComParE feature set, which contains 6,373 static features (i.e., statistical functionals including mean and variance) of low-level descriptor (LLD) contours (i.e., MFCCs) often used in tasks such as recognition of emotion from speech [10].

We further define a generic ASA task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ that consists of a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$ (or a conditional distribution $P(Y|X)$). The goal of this task is to build an effective and robust predictive function $f(\cdot)$ that is capable of learning transformation rules from the feature space

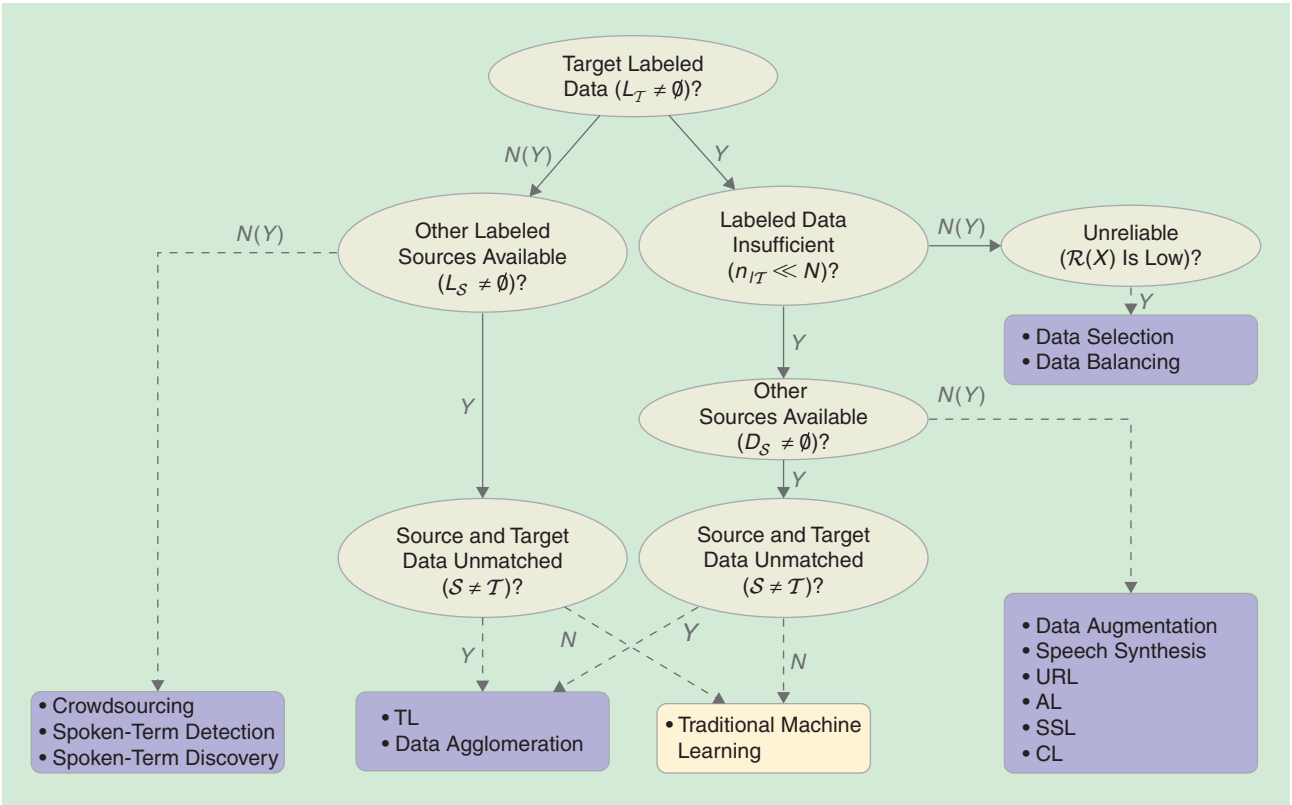


FIGURE 1. A taxonomic overview of the three main data challenges associated with ASA and their potential solutions as discussed in this article. Note that *N*(*Y*) denotes no or yes, which indicates the possible combination of techniques. TL: transfer learning; AL: active learning; SSL: semisupervised learning; CL: cooperative learning; URL: unsupervised representation learning.

$\mathcal{X}$ to the label space $\mathcal{Y}$, i.e., $\mathcal{X} \xrightarrow{f(\cdot)} \mathcal{Y}$. Then, when given a test sample, it maps this feature vector $\mathbf{x}_*$ into a specific label y_*, i.e.,

$$y_* = f(\mathbf{x}_*), \tag{1}$$

where $\mathbf{x}_* \in \mathcal{X}$ and $y_* \in \mathcal{Y}$. As an example, when performing ASR, $y_* \in \mathcal{Y}$ denotes a phoneme or a word; $f(\cdot)$ is then trained to predict a phoneme or a word from, e.g., MFCCs. In speaker identification/verification, $y_* \in \mathcal{Y}$ denotes a speaker identity; the $f(\cdot)$ is trained to predict speaker identity, e.g., from *i*-vectors. Similarly, in speech emotion recognition, $y_* \in \mathcal{Y}$ denotes an emotional state, and $f(\cdot)$ is trained to recognize the emotional state, e.g., from high-dimensional statistical features.

Given a domain $\mathcal{D}$ and a task $\mathcal{F}$, we define D to denote a speech database. As the majority of available pattern recognition approaches are supervised paradigms [the input and expected output for $f(\cdot)$ are provided during training]. A database is normally given by two parts: the feature vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}$ and the corresponding labels $Y = \{y_1, \ldots, y_n\} \in \mathcal{Y}$. However, in real life, the labels y_i are often only partially provided (or not even provided) because of the difficulty of labeling. In this case, we denote the labeled data partition as $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_l}, y_{n_l})\}$ and the unlabeled data partition as $U = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_u}\}$, where n_l and n_u are the total number of labeled and unlabeled instances, respectively. In this sense,

$$D = L \cup U, \tag{2}$$

and $n = n_l + n_u$.

Furthermore, we define the domain for the target task to be the target domain $\mathcal{T}$. The data in this domain might be insufficient for training an effective and robust prediction function $f(\cdot)$. For example, when performing ASR on a low-resource language, $\mathcal{T}$ could be a language such as Assamese, Bengali, Haitian, Lao, Pashto, Tamil, Tagalog, Xitsonga, or Zulu [11]. In this case, we define other domains from which data could be leveraged for the target task as source domains $\mathcal{S}$. For example, for low-resource ASR, one $\mathcal{S}$ could be a high-resource language such as English or Mandarin [11]). According to (2), then

$$D_{\mathcal{T}} = L_{\mathcal{T}} \cup U_{\mathcal{T}}, \tag{3}$$

and

$$D_{\mathcal{S}} = L_{\mathcal{S}} \cup U_{\mathcal{S}}. \tag{4}$$

In this article, we use the term *data* interchangeably with *instance*, *turn*, *record*, *utterance*, *segment*, *sample*, or *example*. Similarly, the term *annotator* is interchangeable with *evaluator*, *transcriber*, *labeler*, or *translator*; and the word *annotation* is used to denote any labeling task, i.e., transcription for ASR or labeling the emotion or other speaker states and traits associated with an utterance.

Data challenges

This section offers a detailed overview into the data sparsity, data unreliability, and unmatched data challenges. Techniques to adequately cope with these challenges will play an essential role in the development of the next generation of reliable ASA systems.

Sparse data challenge

While there is an abundance of raw speech data, the corresponding annotations needed for many ASA tasks are often scarce (i.e., $L_{\mathcal{T}} \neq \emptyset$, but $n_l \ll N$, where $\emptyset$ denotes the empty set and N is a required number), or nonexistent (i.e., $L_{\mathcal{T}} = \emptyset$). For example, outside of speech recognition tasks on a handful of widely used languages (e.g., English and Mandarin), the labels needed to conduct ASR on other languages are particularly scarce (see the Intelligence Advanced Research Projects Activity [IARPA] Babel project [11]). Similarly, most databases available for computational paralinguistics tasks, such as emotion recognition and personality analysis, may contain 5 h of labeled data at most [12], [13], which is insufficient for building highly robust models.

Thanks to the pervasive sensing opportunities offered by smart devices and social media, the gathering of speech data has become a somewhat easier task.

However, thanks to the pervasive sensing opportunities offered by smart devices and social media, the gathering of speech data has become a somewhat easier task. For example, it is reported that some 500 h of video content is being uploaded to the video-sharing website YouTube every minute [14]. Nonetheless, labeling these data demands huge amounts of expert manual labor, which is regarded as being prohibitively expensive and time consuming. Taking speech transcription as an example, it can take up to approximately 6 h to accurately transcribe 1 h of speech at an average price of US$150/h [15], [16]. While a few Internet giants (e.g., Amazon, Google, and Microsoft) have the capability of obtaining many thousands of annotated speech data for ASA tasks, such as speech recognition, these labeled data are, however, rarely made freely available to interested research groups.

If $D_{\mathcal{T}}$ does not contain any labeled data, i.e., $L_{\mathcal{T}} = \emptyset$, a naïve solution is manual annotation. An efficient way to do this is using a crowdsourcing platform, which is an Internet-based system that utilizes a large group of individuals to perform a common service. Alternatively, spoken-term detection/discovery can be considered as a means of detecting predefined patterns in the data or discovering unknown patterns there.

If $D_{\mathcal{T}}$ contains some labeled data, i.e., $L_{\mathcal{T}} \neq \emptyset$, it is then necessary to assess whether or not the available labeled data are sufficient in terms of quantity and diversity to develop a robust model. If the data are found to be insufficient, data augmentation approaches, which seek to enrich the number and variety of existing labeled speech data, might be an appropriate option. A further option is the use of speech synthesis to automatically generate data with predefined labels. If a large scale of unlabeled data are available, i.e., $U_{\mathcal{T}} \neq \emptyset$, alternative solutions could include unsupervised representation learning (URL), semisupervised learning (SSL), active learning (AL), and cooperative learning (CL). These techniques are becoming prominent paradigms to efficiently leverage massive unlabeled data via a small amount of labeled seed data [17].

Unreliable data challenge

This is the scenario in which the total amount of speech data is large, but the data reliability is low. Data sets collected in real-life settings, and even many collected in controlled laboratory settings, are susceptible to a range of problems, such as distortion by environmental noises, recording devices, or interfering speakers [18]. Besides, the associated annotations may be unreliable because of mistakes or high uncertainty among multiple annotators [18]. Furthermore, in many cases, the distribution of collected speech data can be highly unbalanced over the classes of interest. All these factors can give rise to noisy and unreliable data, leading to nontrivial difficulties when training models [18], [19].

Additionally, the reliability of the labeled data should be evaluated in terms of properties such as acoustic quality, annotation certainty, and data balance degree. Poor data quality has frequently shown its detrimental effect on system performance. In this scenario, data selection should be considered for eliminating the noisy, unrelated, and unreliably labeled data or data balancing for balancing the data distribution.

Unmatched data challenge

This is the situation where data from a target domain $\mathcal{T}$ are not sufficient or reliable enough to train a robust model for a task of interest. However, as previously discussed, there are often data from a source domain $\mathcal{S}$ that are easy to obtain and somehow related to the target data. This motivates researchers to explore leveraging source domain data to aid the target ASA task. For example, one of the goals of the IARPA Babel project is to utilize the available and large-scale speech data in, e.g., the English language for speech recognition in low-resource languages. Nevertheless, in many real-world applications, the source and target domains are often highly unmatched in respect to acoustic signal conditions, speakers, tasks, or even recording devices [20]. These mismatches lead to a marked performance degradation of the analysis in such models in real-life settings [20], [21].

Mathematically, the source domain can differ from the target domain (i.e., $\mathcal{S} \neq \mathcal{T}$) in terms of 1) modalities, i.e., $\mathcal{X}_{\mathcal{S}} \neq \mathcal{X}_{\mathcal{T}}$ (this case is considered out of the scope of this article, which is focused only on speech), 2) marginal probability distributions, i.e., $P(X_{\mathcal{S}}) \neq P(X_{\mathcal{T}})$, 3) label spaces, i.e., $\mathcal{Y}_{\mathcal{S}} \neq \mathcal{Y}_{\mathcal{T}}$, and/or 4) conditional probability distributions, i.e.,

$P(Y_S|X_S) \neq P(Y_T|X_T)$. A more in-depth explanation of these discrepancies can be found in [21].

An idealized solution to mitigate these differences is to obtain access to all possible variations by acquiring data on a massive scale. However, it is either practically impossible to anticipate all variations or such data would require exhaustive annotation. In such unmatched scenarios, transfer learning (TL) [21] is regarded to be a highly promising technique to take advantage of the knowledge from the source domain for the target domain.

Finally, it is important to note that all of the aforementioned techniques for each challenge can be performed either individually or jointly. This is illustrated in Figure 1, where possible combinations that can occur are indicated through the use of the N(Y) symbol, which denotes *no* or *yes*. For example, crowdsourcing can be used no matter whether the labeled target data are available or not. Likewise, AL can be executed on either unlabeled target data or unlabeled source data. All of the key techniques mentioned in this section are reviewed in detail in the following sections.

Contributions of this article

The literature shows a few surveys relevant to the topic of this article. Deng et al. [22] offered a comprehensive overview of machine-learning paradigms for speech recognition systems. Wang et al. [20] provided a TL survey for speech and language processing, drawing the conclusion that TL has the potential to overcome the data-mismatch challenge. None of these surveys, however, perform a complete analysis of the sparse, unreliable, and unmatched data challenges or provide a comprehensive overview of the corresponding approaches.

Extending from a previous abstract [12], this article is the first to offer a thorough and in-depth overview of the most prominent and state-of-the-art techniques in this direction, including crowdsourcing for efficient data labeling; spoken-term detection/discovery to facilitate learning when there are no labeled data; data augmentation, speech synthesis, URL, SSL, AL, and CL to enable learning when only a limited amount of labeled resources are available; data selection and balancing techniques to facilitate learning from unreliable or unbalanced resources; and TL and data agglomeration to learn unmatched resources.

Rather than simply enumerating a list of associated papers and techniques, the focus of this article is on the analysis of the various data conditions and on how to better explore data under the different conditions. In doing this, ASA researchers and developers, new and established, can profit from the approaches introduced and discussed for the aforementioned applications.

Efficient data labeling: Crowdsourcing

The most straightforward solution to address a shortage of labeled data is to organize a group of workers (i.e., annotators) to perform the required annotations. By doing this, we create a new or additional labeled data set L_{cs}, and then ASA models can learn from the increased labeled data set $L' = L \cup L_{cs}$. Manual annotation is, however, costly in terms of time and money. Therefore, strategies to reduce these costs are of particular importance.

Crowdsourcing is one method to gather the needed data in a cost-efficient manner. In crowdsourcing, human intelligence tasks (HITs) such as data annotation are distributed via the Internet to a large number of potential workers (annotators). The users perform the tasks for usually low compensation. The assumption behind crowdsourcing is that the use of nonexperts is less onerous and more rapid than the use of experts. Furthermore, the aggregated opinion of many nonexperts has been shown to approach the quality of the opinion offered by comparatively fewer experts [15], [23], [24].

Another emerging trend for crowdsourcing is the gamification of the service, which is used to introduce a sense of fun into what are often simple and recurring tasks.

Popular crowdsourcing platforms include Amazon Mechanical Turk (MTurk), CrowdFlower, and Crowdee. MTurk is likely the most popular crowdsourcing platform for ASA-related tasks. While MTurk provides access to a larger number of potential annotators, it is considered relatively expensive when compared to other platforms [15]. The CrowdFlower platform is steadily increasing its market share. When compared to Mturk, it provides customers with a steady number of contributors and has a higher degree of quality control. An emerging trend, as implemented by Crowdee, involves moving the platform from the web to a mobile platform. Participants associated with this platform have the potential to undertake a task at any time and place.

Another emerging trend for crowdsourcing is the gamification of the service, which is used to introduce a sense of fun into what are often simple and recurring tasks. This is also interesting from an ethical point of view, aiming to improve working conditions of crowd workers. The iHEARu-play platform, for example, offers annotators a chance to perform labeling, or prompted recording tasks, in return for scores and prizes, which are computed on the correctness and workload of their annotations [25].

Generally, the procedure of crowdsourcing speech resources can be broken into four stages. The first step is to define the project parameters, such as an appropriate platform, quality control strategy, budget, and time scale. The second step is to prepare the data. The third step is to distribute tasks. This generally involves splitting the whole task into many small units and then assigning each unit to several annotators. The final step is to aggregate and evaluate the resources (e.g., speech data or annotations).

For speech processing, crowdsourcing has been widely employed for a range of tasks, including speech data collection/acquisition, speech annotation, speech perception, assessment of speech synthesis, and dialog system evaluation [15], [26]. With particular respect to speech annotation, many studies have shown crowdsourcing's benefit in terms

of both increased transcription quality and decreased costs. For example, in [27], the authors proposed a two-stage approach to transcribe speech via a crowdsourcing platform (i.e., microworkers). Specifically, the utterances that were labeled with the lowest agreement level among annotators would be selected for a second-stage translation. In doing this, more than 250,000 utterances (156 h) of spoken dialog from real callers were translated, being of comparable quantity to the same corpora labeled by experts but at considerably less cost. Similar work has been presented for the transcription of meeting data [24], addressing the business name queries from a publicly accessible telephone directory service [16], and labeling the emotional state of speakers [28]. All of these works show that crowdsourcing is a relatively affordable and efficient way to address the task of speech annotation, compared with conventional methods.

Crowdsourcing is a relatively affordable and efficient way to address the task of speech annotation, compared with conventional methods.

Despite the advantages, controlling the quality of the labels is important to ensure they are as reliable as those given by experts. In this regard, quality control measures are required. A range of quality control mechanisms have been proposed in the literature, which can be grouped into one of the following five categories:

1) *Worker filter*: This mechanism evaluates annotation quality through the use of control questions (a question with a restricted answer set) and filters out inappropriate annotations.
2) *Intraworker*: The reliability of an annotator can be evaluated by the consistency of the response to the same question asked multiple times. Alternatively, this could be established by a self-confidence value chosen by the annotator [27].
3) *Interworker*: Normally, a gold standard is calculated via techniques such as majority voting, using responses from a multitude of annotators. The quality of an individual annotator can then be evaluated by calculating the response dissimilarity to the gold standard. This method is, of course, susceptible to the risk that the majority results are wrong.
4) *Gold-standard comparisons*: This is a particular case of the interworker mechanism, where the gold standard is provided by trustworthy experts. This mechanism has been shown to be effective in eliminating intentionally malicious annotators, albeit at the cost of expert intervention [27], [29].
5) *Third-party review*: Here, quality control is carried out by a third party, e.g., another independent crowdsourcing task [30], or by the output of an intelligent system [16], [27]. However, this requires extra quality evaluation or computational costs.

Learning from no labeled resources

This section discusses paradigms suitable for the extreme operating scenario where no labeled data are available, i.e., $L = \emptyset$, and $D = U$. In this scenario, techniques such as spoken-term detection and spoken-term discovery, or related methods of targeted detection of speech-related information and phenomena of interest and according discovery in the sense of novelty detection can be used to identify salient information (i.e., patterns) directly from an unlabeled data set without any manual intervention. The premise of these techniques can be thought of as analogous to infant language acquisition, i.e., the learning of linguistic information from the raw speech of an unknown language during the first few years of an infant's life. The two techniques (i.e., targeted detection and novelty discovery) are distinguished by whether, e.g., spoken terms have been previously identified (spoken-term detection) or not (spoken-term discovery). Next, we focus on terms; however, similar methods can be applied to retrieve speech related to other phenomena of interest.

Spoken-term detection

The goal of spoken-term detection is to retrieve a set of occurrences from a speech repository for given acoustic queries or terms (normally spoken words or phrases). Compared with conventional speech recognition approaches, spoken-term detection offers the capability to detect corresponding patterns from speech in the absence of any text information.

The predominant spoken-term detection methods involve template-based acoustic models and typically rely on dynamic time warping (DTW) [31]. Specifically, they search for the predefined terms in a lattice. In a no-labeled-resource scenario, DTW has been shown to be an effective way to find the matched patterns [31]. Nevertheless, DTW alignment requires substantial computational resources to compare segments [32], [33]. Tackling this runtime-scalability problem is an active and ongoing research direction [32], [34]. Key approaches proposed in the literature include information retrieval-based DTW [35]. This approach first estimates the regions of an utterance that are more likely to contain the spoken query and then uses a standard DTW to find the exact start and end times of each pattern. This approach was further extended in [34] via the introduction of a hierarchical k-means clustering, contributing to a substantial speedup when compared with classic DTW.

An alternative approach is to embed the arbitrary-length segments into fixed-dimensional spaces [32]. This technique greatly reduces the computational load without any performance compromise. Following this idea, the novel framework of audio Word2Vec was recently proposed [36]. Audio Word2Vec uses a sequence-to-sequence autoencoder [a neural network (NN) commonly used as an unsupervised learning algorithm; for more details, see the "Deep Belief Networks and Stacked Autoencoders" section] to represent any arbitrary-length audio segment as a fixed-length vector. This framework was determined to outperform conventional DTW-based approaches at substantially lower computational requirements [36].

Spoken-term discovery

Spoken-term discovery, also known as *spoken-term indexing*, is the task of searching potentially large, untranscribed speech collections for recurring words and phrases without using any language-specific resources other than the collection itself [37]. Specifically, spoken-term discovery differs from spoken-term detection in that spoken-term discovery systems automatically find an inventory of lexical units (words or phrases) without being given any user-specific terms. Furthermore, spoken-term discovery is distinct from conventional ASR systems, where a lexicon is always specified.

Typically, spoken-term discovery consists of three steps [13]: 1) pairwise matching, 2) clustering, and 3) parsing. The aim of pairwise matching is to identify pairs of segments, taken from unique continuous spoken utterances, that have high acoustic similarity. Similar to spoken-term detection, the dominant techniques in this step are based on DTW.

The discovered segments are then clustered into classes (indices) that correspond to a set of likely words and phrases present in the data. Typically, an abstract adjacency graph [31] is used to represent the relationship between all of the segmented pairs. The nodes of this graph correspond to the locations in time of the segments, and its edges correspond to the measures of similarity between those time indexes. A predefined threshold is then applied to the edge weights, which results in clusters of highly connected nodes. While the edge thresholding is regarded as the de facto clustering method for spoken-term discovery, there is a range of fast and efficient algorithms for automatic graphic clustering that could be applied. For example, the work in [31] utilized the Newman algorithm, which first removes all edges and then merges potential groups together in a greedy fashion by adding edges back to the graph.

Finally, the discovered speech segments are used to parse the utterances. The identification of the segment (term) boundaries is challenging; the alignment segments are often overlapping in a particular node, and the ending times of their respective time intervals can differ. A straightforward solution for this issue is to calculate the average start and ending times for all of the alignment segments belonging to one node [31].

While considerable advances have been made for fully unsupervised speech processing, the majority of studies are limited to small-size data sets. Studies have shown that performance is dramatically degraded when facing a large data set [26] or a large variety of speakers [38]. However, this approach is still quite attractive for many low-resource ASA tasks, e.g., early language acquisition.

Learning from limited labeled resources

Rather than starting with a completely unlabeled data set, we are often in the better situation of having a limited number of labeled resources, i.e., some few and expensive labeled speech data exist $L \neq \emptyset$, while $n_l \ll N$, where N denotes an opportune number of annotations. In this scenario, a range of other techniques besides the aforementioned no-labeled-resource methods can be utilized. These are generally implemented in one of two ways: 1) increasing the size and diversity of the existing labeled data by means of manually modifying the speech variations (i.e., data augmentation) or artificially generating new speech with predefined labels (i.e., speech synthesis) or 2) the efficient leveraging of information gained from big unlabeled data, through a priori knowledge of the labeled data. Typical techniques here include URL, SSL, AL, and CL. In the following text, each of these techniques is discussed in detail, with key contributions from the literature summarized in Table 1.

Data augmentation artificially generates more data by transforming existing speech samples using certain transformations that preserve the original class labels and speech content.

Data augmentation

Data augmentation artificially generates more data by transforming existing speech samples using certain transformations that preserve the original class labels and speech content. By taking this approach, an augmented data set L_{aug} is obtained from the original data set L, i.e., $L_{\text{aug}} = AUG(L)$, which is then added to an updated labeled data set $L' = L \cup L_{\text{aug}}$. The popularity of data augmentation is indeed highly relevant to the ongoing development of deep learning, the success of which strongly depends on having large amounts of training data. Many studies have reported that training on data of limited quantity and variety leads to a failure of deep-learning systems owing to factors such as overfitting [6].

Variations in speech data are strongly influenced by numerous factors, such as the speaker's age, gender, and cultural background, and even the content of the background noise. Data augmentation techniques, through a series of transformations (perturbations), allow us to artificially increase both the quantity and variations present in some training data, consequently improving the generalizability of the classifiers trained on this data. Conventional data augmentation approaches mainly involve artificially adding noise of various types, including convolutional noise, and levels to the original training speech for training a noise-robust acoustic model in multiple acoustic conditions [39].

Recently, research efforts have focused on using more complex perturbation approaches, such as vocal tract length perturbation (VTLP) [40], or stochastic feature mapping (SFM) [41]. In VTLP, an alternate replica of an utterance is created by distorting its spectrum [40]. First, Mel-filter banks are applied over the spectrum. Then, the center frequencies (f) of all of the filter banks are mapped to new frequencies (f') by employing a warping procedure:

$$f' = f \cdot \phi(\alpha), \tag{5}$$

Table 1. Selected data-exploitation studies on the limited labeled speech resource.

Publications	Types	Approaches	Models	Applications	Databases and Languages
Weng et al. 2014 [39]	DAU	Adding noise	Recurrent DNN	ASR	WSJ0 (En)
Amodei et al. 2015 [7]	DAU	Adding noise	CNN, DNN, CTC	ASR	WSJ0 (En), Switchboard (En), Fisher (En), Baidu (En, Ma), LibrisSpeech (En)
Jaitly and Hinton 2013 [40]	DAU	VTLP	DNN, CNN	ASR	TIMIT (En)
Cui et al. 2015 [41]	DAU	VTLP, SFM	DNN, CNN	ASR, KWS	IARPA Babel program (As, Ha)
Tüske et al. 2014 [42]	DAU	VTLP	BN-MLP	ASR, KWS	IARPA Babel program (five lang.)
Ko et al. 2015 [43]	DAU	Tempo-/speed based	Time Delay NN	ASR	Switchboard (En), Gale database (Ma), LibriSpeech (En), Tedlium (En)
Peddinti et al. 2015 [44]	DAU	Volume based	Time Delay NN	ASR	Switchboard (En)
Milde and Biemann 2015 [45]	DAU	Pitch based	CNN	Eating condition classification	iHEARu-EAT corpus (En)
Schuller et al. 2012 [46]	SS	Waveform-based	SVM	ER	Two synthesized + eight human corpora
Gales et al. 2009 [47]	SS	Parameter-based	SVM, HMM	ASR	WSJ Corpus (En)
Dahl et al. 2012 [51]	URL	DBNs	DBNs	ASR	Business Search Dataset (En)
Seide et al. 2011 [64]	URL	DBNs	DBNs	ASR	Switchboard-I (En)
Deng et al. 2010 [54]	URL	SAEs and DBNs	SAEs and DBNs	Speech coding	TIMIT (En)
Mohamed et al. 2012 [65]	URL	DBNs	DBNs	ASR	TIMIT (En)
Lei et al. 2014 [66]	URL	DNNs	DNNs	Speaker recognition	NIST SRE'12 (En)
Liu et al. 2014 [67]	URL	DBNs	DBNs	Speaker identification	NIST 2005 SRE (En)
Stuhlsatz 2011 [68]	URL	DNNs	DNNs	ER	Nine emotional corpora
Sánchez-Gutiérrez et al. 2014 [69]	URL	DBNs	DBNs	ER	Spanish emotional speech database (Sp)
Kim et al. 2013 [70]	URL	DBNs	DBNs	Audiovisual ER	IEMOCAP (En)
Hau and Chen 2011 [57]	URL	Deep CNNs	Deep CNNs	Speaker/gender identification Phone classification	TIMIT (En)
Lee et al. 2009 [58]	URL	Convolutional DBNs	Convolutional DBNs	Speaker/gender identification Phone/music classification	TIMIT (En), music data
Kemp and Waibel 1999 [71]	SSL	Self-training	GMM–HMM	ASR	View4You broadcast news database (Ge)
Wessel and Ney 2005 [72]	SSL	Self-training	HMM	ASR	BROADCAST NEWS96/7 corpora (En)
Fazakis et al. 2015 [73]	SSL	Self-training	NB, SVM, LR	Speaker identification	CHAINS Corpus (En)
Hsiao et al. 2013 [74]	SSL	Self-training	MLP	KWS	IARPA Babel Program (Tu, Vi)
Thomas et al. 2013 [75]	SSL	Self-training	DNN	ASR	Callhome Corpora (En, Ge, Sp)
Zhang et al. 2013 [76]	SSL	Cotraining	SVM	Emotion/sleeping/ age/gender classification	Six emotional corpora
Cui et al. 2012 [77]	SSL	Multiview learning	RDT, HMM	ASR	Broadcast News corpus (En)
Liu and Kirchhoff 2016 [78]	SSL	Graph-based learning	DNN	ASR	Switchboard (En), DARPA RM (En)
Riccardi and Hakkani-Tür 2005 [79]	AL	Uncertainty sampling	HMM	ASR	"How May I Help You?" database (En)
Varadarajan et al. 2009 [80]	AL	Uncertainty sampling	HMM	ASR	Directory assistance data (En)
Fraga-Silva et al. 2015 [81]	AL	Uncertainty sampling	GMM–HMM	ASR, KWS	IARPA Babel Program (six languages)

(*continued*)

Table 1. Selected data-exploitation studies on the limited labeled speech resource. (*continued*)

Publications	Types	Approaches	Models	Applications	Databases and Languages
Hamanaka et al. 2010 [82]	AL	Query by committee	GMM–HMM	ASR	Corpus of Spontaneous Japanese (Ja)
Zhang and Schuller 2012 [83]	AL	Meta query	SVM	ER	FAU AEC (Ge)
Zhang et al. 2015 [84]	AL	Meta query	SVM	ER	FAU AEC (Ge)
Riccardi and Hakkani-Tür 2003 [85]	CL	Confidence score	HMM	ASR	"How May I Help You?" database (En)
Yu et al. 2010 [86]	CL	Confidence score	HMM	ASR	Broadcast Conv. and News corpora (Ma)
Zhang et al. 2015 [17]	CL	Confidence score	SVM	ER	FAU AEC (Ge), SUSAS (En)
Yu et al. 2010 [87]	CL	Global-entropy based	HMM	ASR	Directory assistance data (En)

BN: Bayes network; CTC: connectionist temporal classification; NB: naive Bayes; LR: logistic regression; RDT: randomized decision making; DAU: data augmentation; SS: speech synthesis; ER: emotion recognition; As/Da/En/Fr/Ge/Ha/Ja/Ma/Sp/Tu/Vi/Xi/Zu: Assamese/Danish/English/French/German/Haitian Creole/Japanese/Mandarin/Spanish/Turkish/Vietnamese/Xitsonga/Zulu.

where α, the wrapping factor, is randomly chosen from [0.9, 1.1]. The results presented in [40] indicate that, in terms of the phone error rate, deep networks trained on a VTLP-augmented version of a small database can outperform the deep networks trained on the original data set. Based on that work, a deterministic perturbation (i.e., α changes in the range of warping factors with a fixed step) rather than a random perturbation was proposed and investigated [42].

SFM, inspired by voice conversion paradigms, seeks to utilize the acoustic-feature-space relationship among speakers when augmenting a data set [41]. Specifically, it augments training utterances by statistically converting one speaker's speech data to another's using

$$\mathbf{x}' = \mathbf{x} \cdot \mathcal{M}, \tag{6}$$

where $\mathcal{M}$ is a transformation matrix of the feature spaces between two speakers. The experimental results given in [41] show that SFM offers improved performance over VTLP on both ASR and keyword spotting (KWS) tasks.

Other data augmentation approaches include tempo-based, speed-based, and volume-based perturbations [43]. Tempo-based perturbation modifies the speech tempo while retaining the pitch and the spectral envelope. Speed-based perturbation varies the speech speed by resampling, whereas volume-based perturbation changes the amplitude of signals.

While data augmentation approaches have frequently been effective in ASR tasks [7], [44], this has not proved to be as much the case in other ASA tasks, particularly in computational paralinguistics [45]. A potential reason for this might be that the detection of speaker states and traits (e.g., emotion, age, and gender) is more sensitive to changes in speech variation. Therefore, training on inappropriately transformed speech would lead to a worse model. Emotion, for example, is known to be related to the speech tempo; speech with faster tempo is inclined to be recognized as higher arousal in emotion recognition, so changing the associated speech tempo from fast to slow would potentially lead to badly labeled training data.

Continued research efforts being undertaken to distinguish features that are task specific or task invariant could help facilitate the application of data augmentation to other speech analysis tasks. In addition, most recent applications of data augmentation are performed for deep learning [7]. The effectiveness of these techniques on shallow discriminative or generative models is yet to be established.

Speech synthesis

Similar to data augmentation, the speech synthesis approach aims to synthesize additional labeled data, i.e., $L_{\text{syn}} = SS(L)$, such that the new labeled data set L' is updated by $L' = L \cup L_{\text{syn}}$. Theoretically, speech synthesis can produce an infinite amount of labeled data via altering speech content or modifying the parameters of a speech synthesizer. However, as the parameters of the synthesizers have a limited range, the simulated speech data often face the problem of limited variations. This can consequently result in the overfitting issue when training models. Combining the synthesized speech data with natural instances has been shown to help minimize this overfitting issue [46]. For emotion recognition in speech, it has been shown that systems trained on synthesized speech (the test data was natural speech) can deliver competitive performance when compared to equivalent systems trained on natural speech [46]. In this article, two synthesizers rendering emotional speech—Emofilt and Mbrola—were utilized to artificially generate speech colored with predefined emotions [46].

Rather than directly synthesizing waveforms, an alternative is generating parameterized speech that can be used directly for training a discriminative classifier. Gales et al. [47] used a hidden Markov model (HMM)-based statistical synthesis to generate missing words in a training set, when building word-based support vector machines (SVMs) for ASR. The results presented indicate that this HMM-based synthesis approach was able to yield gains over the baseline. Inspired by the success of deep learning, an emerging research trend is to use NNs rather than HMMs to generate speech samples [48], which may also mature in terms of the variation of synthesized speaker states and traits.

Unsupervised representation learning

In contrast to data augmentation and speech synthesis, URL techniques attempt to leverage massive unlabeled data, rather than sparsely labeled data. URL is closely related to the pretraining process of deep learning, which aims to learn the underlying representations $\mathbf{x}'$ embedded in speech signals via multiple unsupervised transformations, i.e., $\mathbf{x}' \leftarrow URL(\mathbf{x})$, where $\mathbf{x} \in D = L \cup U$. To train a recognition model for a specific task, the pretrained model is then updated in a supervised manner via a small amount of labeled data. This step is generally referred to as *fine-tuning* or *discriminative learning*.

A typical model structure for URL is often composed of multiple processing layers of NNs for linear and nonlinear transformations (Figure 2). To efficiently train such a DNN, Hinton and Salakhutdinov [49] introduced a greedy layer-wise unsupervised algorithm to initialize multiple-layer feedforward NNs. Since then, this training algorithm has been frequently shown to have a powerful capability to capture representative features via massive unlabeled data, and has obtained tremendous success in a variety of applications, particularly in the context of ASA [7], [50], [51]. The remainder of this section introduces several of the most important deep architectures for URL, including deep belief networks (DBNs), stacked autoencoders (SAEs), convolutional NNs (CNNs), and recurrent NNs (RNNs).

In contrast to data augmentation and speech synthesis, URL techniques attempt to leverage massive unlabeled data, rather than sparsely labeled data.

Deep belief networks and stacked autoencoders

Two of the most established deep-learning architectures are DBNs and SAEs. These topologies are formed by stacking multiple layers of restricted Boltzmann machines (RBMs) or feedforward autoencoders, respectively. The unsupervised pretraining of these architectures is done one layer at a time.

For SAEs, each layer is trained with an encoder $h(\cdot)$ and a decoder $g(\cdot)$ by minimizing the reconstruction error at its input $\mathbf{x}$:

$$g(h(\mathbf{x})) \approx \mathbf{x}. \tag{7}$$

The output of the encoder $h(\mathbf{x})$ forms an alternative representation of the input $\mathbf{x}$ and is fed into the successive layer as input. This procedure is repeated layer-by-layer until all predefined layers are initialized. The training of the stacked layers in this manner allows a deep network to incrementally learn a more robust representation when compared to training the whole network, in ensemble, from a random initialization of weights. For further insights into the advantages of pretraining with autoencoders and RBMs, see [52]. This observation is particularly true for stacked denoising autoencoders [53], extensions of SAEs where the initial input $\mathbf{x}$ is partially corrupted into another version $\tilde{\mathbf{x}}$ by means of stochastic mapping, i.e., $\tilde{\mathbf{x}} \sim q_d(\tilde{\mathbf{x}} \mid \mathbf{x})$. The robustness of the high-level representations formed using this technique is improved when compared to the aforementioned SAE [53].

An early attempt at applying deep-learning technologies to learn speech representations was proposed by Deng et al. [54], where the authors utilized DBNs and deep SAEs to compress (represent) speech directly from spectrograms. When compared with the traditional compression approach of vector quantization, this technique showed a much lower log-spectral distortion over the entire frequency range of wide-band speech. Expanding on the work of this article, DBNs have been extensively tested as an acoustic modeling paradigm for speech recognition and have shown encouraging performance in comparison with the conventional Gaussian mixture model (GMM) and HMM-based acoustic models for ASR [51]. For an overview of deep URL models for ASR and the corresponding performance gains, the reader is referred to both [6] and [50]. Inspired by these achievements, deep URL techniques have started to become the dominant approach in almost all areas of speech processing.

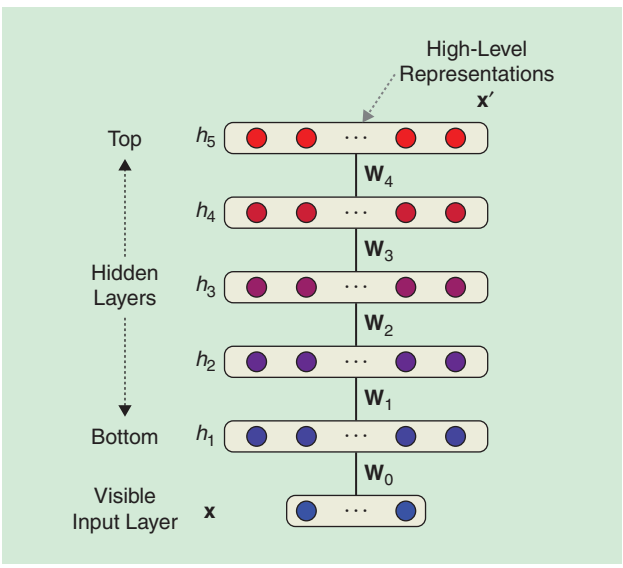


FIGURE 2. An illustration of typical deep URL. Usually, each layer of the network is individually trained in an unsupervised manner; this allows the network to incrementally learn a more robust representation than the one learned by training the network as a whole.

Convolutional neural network

Another deep architecture currently exciting great interest is the CNN [55], [56]. CNNs are a biologically inspired variant of the multilayer perception (MLP) originally developed for visual perception tasks [55]. Typically, they consist of one or more convolutional layers (often with a subsampling layer), followed by one or more fully connected layers.

CNNs are normally trained in a supervised manner. However, unsupervised training approaches are gaining in popularity. Inspired by the unsupervised learning algorithm for DBNs, Hau and Chen [57] constructed a deep architecture using a CNN trained in an unsupervised manner as an alternative

building block to retrieve effective hierarchical speech representations. Specifically, the authors utilized an unsupervised predictive sparse decomposition algorithm to train the weights of the encoder and decoder [57].

Furthermore, a combination structure of CNNs and DBNs was proposed in [58], in which the authors constructed convolutional DBNs (CDBNs) with convolutional RBMs (CRBMs) as the building blocks. The CRBM is an extension of the conventional RBM to a convolutional setting. The weights between the hidden units and the visible units are shared among all locations in the hidden layers [59]. By leveraging a large amount of unlabeled data, the authors demonstrated that the learned hierarchical CDBN representations are competitive with conventional features (e.g., MFCCs) when evaluated across multiple audio classification tasks.

Long short-term memory recurrent neural networks

Unlike the aforementioned NN architectures, RNNs allow cyclical connections, which consequently endow the network with the capability of accessing previously processed information (i.e., context sensitivity). An advanced version of this paradigm, the long short-term memory (LSTM)–RNN [60], has recently attracted a large amount of attention. An LSTM unit contains one input, one output, and one forget gate to control the memory cell, which enable it to store and access information over a long temporal range. Therefore, the LSTM–RNN combination has a powerful capability for sequence learning.

In utilizing the advantages associated with LSTM–RNNs, Srivastava et al. [61] recently proposed and explored an unsupervised sequence-to-sequence learning paradigm where the LSTM–RNNs are constructed as an encoder–decoder. By doing this, the system efficiently learns the underlying representations of video sequences for future frame prediction or sequence reconstruction. This model has been further investigated by Chung et al. [36] for audio segment representations, where the authors demonstrated its effectiveness for spoken-term detection when compared with classic DTW. More recently, the gated recurrent unit has emerged as a computationally simpler alternative to the LSTM unit [62].

Overall, deep unsupervised learning paradigms have seemingly great potential for learning useful representations of large-scale unlabeled speech data. Nevertheless, in most cases, it is necessary to implement additional supervised training, such as fine-tuning, to ameliorate the system for a specific application [51], [63]; therefore, a small amount of labeled data is often additionally required to produce state-of-the-art performance.

Semisupervised learning

Unlike URL, which aims to distill representative features from unlabeled speech, SSL is designed to enhance recognition models. Given a seed set of labeled data, SSL exploits information from a large set of unlabeled data in an efficient manner with minimal intervention from human annotators. SSL methods are generally distinguished as being conducted in either an inductive or transductive manner [88]. The primary discrepancy between them lies in whether the distribution information of the unlabeled data is utilized for their own prediction.

Inductive approaches require the construction of a classification model f based on a priori knowledge of labeled data. The predictive model f is then used for predicting the unlabeled data, no matter whether they are presented in an online (afterward) or offline (beforehand) manner. Hence, inductive approaches are also known as a *supervised learning + additional unlabeled data* paradigm. Mathematically, this can be expressed as

$$\{(\mathbf{x}^l, y^l), l = 1, \ldots, n_l\} \mapsto f,\ f \mapsto \{y^u, u = 1, \ldots, n_u\}. \tag{8}$$

Once the automatically predicted annotations have been obtained from the unlabeled data set L^*_{ssl}, the labeled training set is updated, i.e., $L' = L \cup L^*_{\text{ssl}}$.

In contrast, transductive approaches do not need to prebuild a classification model f but instead perform predictions directly on the unlabeled data by exploiting the joint probability distributions of labeled and unlabeled data sets. In this technique, the unlabeled data set should be available beforehand. When new samples arrive, the transductive algorithms have to be rerun, which consequently increases the computational load. Hence, the transductive approaches are also referred to as the *unsupervised learning + additional labeled data* paradigm. That is,

$$\{(\mathbf{x}^l, y^l), l = 1, \ldots, n_l\} \cup \{\mathbf{x}^u, u = 1, \ldots, n_u\} \mapsto \{y^u, u = 1, \ldots, n_u\}. \tag{9}$$

Note that both the inductive and transductive approaches can be jointly deployed, as in transductive SVMs in which unlabeled data are also considered when determining the hyperplane [89].

The ASA literature is dominated by inductive SSL approaches. This is possibly due to inductive approaches being more flexible to the availability format of unlabeled data (i.e., online or offline). Among the inductive SSL approaches proposed, self-training (i.e., self-teaching) is arguably the most representative and has been widely and efficiently used for ASR [71], [72], emotion recognition [90], and speaker identification [73]. (In the context of ASR, SSL is often referred to as *unsupervised learning* or *unsupervised training*.)

A typical self-training paradigm is based on prediction uncertainty. That is, those samples $\{\mathbf{x}'^u_i\}$ recognized with high confidence C are picked up and combined into a selected subset S, and those $\{\mathbf{x}^u_j\}$ with low confidence remain in the unlabeled data set U:

$$\underset{\forall \mathbf{x}'^u \in S}{C(\mathbf{x}'^u)} \geq \underset{\forall \mathbf{x}^u \in U \setminus S}{C(\mathbf{x}^u)}. \tag{10}$$

The selected data set $\mathcal{S}$ (together with their pseudolabels) is then combined with the initial training set L to form a new

data set $(L' = L \cup L^*_{\text{ssl}})$, which is sequentially employed to refine the previous model and retest the remaining unlabeled data. This process is repeated several times to incrementally upgrade the initial model.

Self-training is simple and can be easily applied to an existing model. However, it is open to the risk of error accumulation, which is introduced by the selection of misclassified data in early learning iterations. Commonly used techniques to mitigate such a detrimental effect include 1) using an additional development partition to determine the stopping point of learning, 2) using generalized expectation maximization to assign weights to the automatically labeled data based on the prediction confidence [74], and 3) retesting previously selected data for subsequent reevaluations and selections, such that the mislabeled data in previous iterations are possibly corrected in future iterations with an improved model [91].

Another commonly used inductive SSL paradigm in ASA is cotraining. Compared with self-training, cotraining attempts to exploit the mutual information between two learners (trained on different views or feature domains $\mathcal{X}_1$ and $\mathcal{X}_2$). That is, each learner uses its own predictions to teach not only itself, but also the other learner [92].

Successful cotraining relies on two assumptions: sufficiency and conditional independence [92]. Sufficiency infers that each view is sufficient for classification on its own, i.e., the two hypotheses $f_1: \mathcal{X}_1 \mapsto \mathcal{Y}$ and $f_2: \mathcal{X}_2 \mapsto \mathcal{Y}$ are good enough for recognition. Conditional independence denotes that the views are conditionally independent, given the class label, i.e., $P(y_i|\mathbf{x}) \leftarrow P(y_i|\mathbf{x}_1)P(y_i|\mathbf{x}_2)$. Although these two assumptions are restrictive, the work presented in [76] shows the capability of cotraining for retrieving emotional information in unlabeled data via separating the acoustic feature set into two pseudo views (i.e., not completely conditional independence) in the speech domain. Similar verification of cotraining has also been reported for other computational paralinguistics tasks [76]. Additionally, a more general framework called *multiview learning* requires less restriction in terms of conditional independence than cotraining and has been successfully applied in speech recognition by using several types of acoustic features and randomized decision trees [77].

More recently, SSL research in ASA has started to explore the advantages of deep-learning techniques [75], [93]. A typical implementation is ASR for a low-resource language [75], [93]. First, an initial DNN is trained in an unsupervised manner using multilingual data to learn the generalized representation of speech. Next, this model is fine-tuned as a seed model by using limited amounts of monolingual data from the low-resource language. The seed model is then employed to decode the untranscribed utterances, with the predicted hypotheses being regarded as the training transcripts for the next iteration. Various discriminative criteria (e.g., maximum mutual information or minimum cross entropy) can be adopted to obtain the prediction confidence scores for each frame, word, or utterance [75], [93]. Similar to traditional self-training and cotraining, the data (i.e., frame, word, or utterance) predicted with high confidence are assumed to be of high quality and are then incorporated to update the initial DNN or GMM–HMM acoustic model.

Apart from the inductive approaches, a graph-based transductive approach can also be integrated into DNN-based speech recognition systems at either a late or early stage [78]. For the late-stage integration, a graph is first constructed over the labeled and unlabeled data sets, where the node represents a data instance and the edge indicates the similarity between a data instance pair. Then, using a graphic-based learning algorithm, a new set of posterior distributions for each instance of unlabeled data is produced. After that, the posteriors are converted into a graph likelihood and are integrated with the original acoustic scores given by the DNN for a subsequent rescoring of the unlabeled data [78]. A major drawback of this late integration approach is a substantially increased computational cost, as the graph has to be reconstructed after each learning iteration. To overcome this problem, an early-stage integration algorithm has been proposed [78]. This algorithm employs a graph embedding approach in which the data in the graph is transformed into a compact feature vector, which is then used as additional input for the DNN.

Active learning

Similar to SSL, AL attempts to improve recognition models by exploring unlabeled data. However, unlike SSL, which performs automatic machine (model) annotation, the focus of AL approaches is to efficiently select the most informative data S in the unlabeled collection U for manual annotation. Partly because of the growing amounts of data to be handled and the popularity of crowdsourcing (see the "Efficient Data Labeling: Crowdsourcing" section), AL strategies for ASA are currently more important than ever.

One of the central goals of AL is to determine the informativeness of unlabeled data, a process known as *query strategy*. The following sections briefly review the most commonly used strategies with relevance to ASA, which include the uncertainty sampling, query by committee, and metaquery strategies.

Uncertainty sampling

This strategy uses confidence measures as a criterion to select the most informative data. The basic idea is to use a pretrained model (an active learner) to determine the uncertainty of predictions for a specific ASA task. The instances with the least certain predictions are then sent to an oracle (a human) for the annotation.

Formally, the selected data can be expressed as

$$\mathbf{x}' = \arg\min_{\mathbf{x} \in U} Q_c(\mathbf{x}; \theta), \tag{11}$$

where θ indicates the model parameters trained on the labeled data set L and Q_c denotes the confidence measure function.

When using a probability model (e.g., Bayesian networks), this function is usually estimated using either the posterior

probability, the probability margin between the two most likely class labels, or the entropy of prediction [94]. In the context of speech recognition, word posterior probabilities or the HMM-state entropy are frequently used as confidence measures [79], [81]. When using a nonprobability model (e.g., an SVM), similar measures can be constructed from discriminant functions. Considering the SVM as an example, pseudoprobabilistic values can be transformed from the output distances from the SVM hyperplane (see [17] for more details). The effectiveness of this approach has been extensively assessed for emotion recognition from speech [83].

Despite the reported performance improvement, many studies have found that uncertainty-based AL is inclined toward selecting noise and garbage data (i.e., outliers from the main data distribution) for human labeling. This issue occurs even more frequently when using AL to annotate data collected in the wild, i.e., not under controlled laboratory conditions, where environmental noises severely distort the speech, and many unexpected words are potentially uttered. Labeling these outliers is usually difficult and time consuming [95]. Furthermore, these data often offer little information on the overall system performance [17], [95]. A straightforward solution to address this outlier problem is to raise the threshold of a confidence score. For example, the authors of [17] used a median uncertainty strategy instead of the least certainty one for actively selecting spontaneously emotional utterances, which delivered a positive performance improvement.

Sampling by uncertainty and density (SUD) is a more sophisticated method that was introduced for ASR in [96]. In this approach, unlabeled instances that are both near the decision boundary and very close to other examples are assumed to be more important than those that are isolated (i.e., likely to be outliers). Hence, SUD considers not only the most informative data in terms of uncertainty but also the most representative data in terms of density. That is, those data predicted with least certainty and distributed in a low-density area are ignored.

A similar idea was proposed in [80], where the global criterion was used in ASR to maximize the expected lattice entropy reduction over all nontranscribed data. Specifically, it first measures the entropy among the lattices generated by decoding unlabeled utterances. It then estimates the expected entropy reduction over the whole data set for each given utterance, and selects the utterances that should deliver the highest entropy reduction for human labeling. After that, the transcribed utterances can be weighted according to the number of similar utterances in the whole data set to achieve better performance for speech recognition. This algorithm is also analogous to the error-rate reduction strategy introduced in [95].

Query by committee

This strategy uses a committee (group) of weak models (learners), denoted by $\Theta = \{\theta_1, \ldots, \theta_k\}$, to select unlabeled data by the principle of maximal disagreement among these models [97]. Mathematically, this can be expressed as:

$$\mathbf{x}' = \arg\max_{\mathbf{x} \in U} Q_d(\mathbf{x}; \Theta). \tag{12}$$

The two key problems in committee-based approaches are 1) constructing a committee Θ that represents competing hypotheses and 2) defining a disagreement measurement Q_d. To alleviate the first problem, the models are usually built by employing multiple different classifiers (e.g., HMMs, SVMs, and RNNs) with the same training data, or by splitting the training data or features into partitions for training several different versions of the same type of classifier, or by a combination thereof. For the second problem, the commonly used disagreement measures are vote entropy and Kullback–Leibler divergence (see [94] for more details). In speech recognition, this strategy has been applied to both acoustic and language models, resulting in a significant data annotation reduction while achieving the same word accuracy [82].

Meta query strategies

One often deals with imbalance across classes of interest in the data. As an example, for emotion recognition, the emotional speech of interest usually appears sparsely within a data set, while the less interesting nonemotional speech often appears at a much higher frequency. In this scenario, an initial coarse model can be used to first decide which data are of interest by distinguishing between neutral and emotional speech. A subsequent finer model can be then used to recognize different emotions or respective other classes in other tasks in the selected emotional speech data. An example of such an approach is the sparse-tracking query strategy [83]. It tracks only sparse (emotional) instances, via iterative retraining and labeling, using a novelty detection paradigm.

One issue when analyzing subjective speaker states and traits (e.g., emotion and personality) is the requirement of multiple annotations per sample to obtain a reliable gold standard, which linearly increases the annotation workload. Recently, dynamic active query strategies have been shown to be successful in overcoming this issue [84]. These approaches, e.g., sequentially query human annotators to label a specific instance up to the achievement of a predefined agreement level (i.e., a certain number of votes for a specific class). The general idea is to learn and exploit the varying reliability of raters to discern whom to best trust and when. The results presented indicate that this approach can contribute to a meaningful reduction of annotation effort [84].

Cooperative learning

As discussed previously, SSL techniques can perform annotation work from machines with a bare minimum of human intervention. However, the performance of SSL is hampered by the issue of potential error accumulation [94]. Alternatively, AL techniques have the potential to achieve higher accuracy with fewer training labels by actively selecting the data it can learn the most from. However, AL still requires a considerable amount of human intervention.

To take advantage of the best of both approaches, it is plausible to jointly conduct AL and SSL in a unified CL framework [17]. A general CL flowchart is illustrated in Figure 3. CL allows the sharing of the labeling effort between human and

machine oracles, while being able to mitigate the limitations of SSL and AL. This is achieved by successively fusing the data subset selected by the AL (L_{al}) and the one selected by SSL (L^*_{ssl}) into the original training set in an iterative fashion. In this case, the labeled data set L' is continuously updated by $L' = L \cup L_{\text{al}} \cup L^*_{\text{ssl}}$. To minimize the effects relating to error accumulation, AL is often conducted before SSL.

Early studies of CL mainly focused on text classification. McCallum and Nigam were the first to investigate the idea of integrating the query by committee-based AL and the expectation maximization-based SSL for text classification [98]. Later, motivated by the success of cotraining (see the "Semisupervised Learning" section), a similar idea of jointly using multiple views was taken into account, contributing to the new CL algorithm of coexpectation-maximization testing [99].

For speech processing, the first CL efforts were undertaken by Riccardi and Hakkani-Tür [85] for ASR. This approach assigned confidence scores to transcribed utterances based on the lattice output, from which the utterances were determined to be manually or automatically labeled. A similar idea was also investigated by Yu et al. [86] for speech recognition. In this approach, the data recognized with high confidence are translated automatically by machine, while the ones recognized at a low confidence are selected and translated manually. Similar to the uncertainty-based AL, this uncertainty-based CL is as well inclined to choose noise and garbage utterances that typically have low confidence scores.

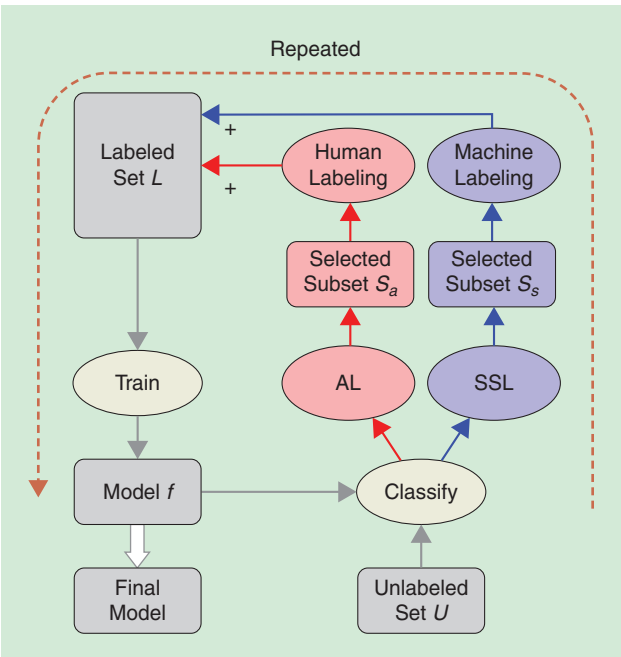


FIGURE 3. A general overview of a CL framework that aims to take advantage of both AL and SSL.

Motivated by the success of the global entropy reduction maximization criterion [80] for AL (see the "Active Learning" section), Yu et al. [87] extended the work of [80] by integrating this approach with SSL. The results presented indicate that this technique achieves a notable performance increase when compared to the uncertainty-based CL approaches for speech recognition. Besides, Zhang et al. [17] recently combined SSL with a median uncertainty-based AL for emotion recognition, which efficiently helps to avoid choosing garbage data as well. Furthermore, in the same article, multiview CL (i.e., where two views are used for both AL and SSL) was exemplified and demonstrated to achieve better performance than the single-view CL [17].

CL is indeed a productive, highly efficient way to exploit unlabeled speech data to enhance the performance of preexisting models while minimizing human work.

Experimental results obtained in the aforementioned studies indicate that, when compared to SSL and AL, CL is indeed a productive, highly efficient way to exploit unlabeled speech data to enhance the performance of preexisting models while minimizing human work. Moreover, its potential is expected to be further evoked when implemented with a crowdsourcing platform (see the "Efficient Data Labeling: Crowdsourcing" section and/or, incorporated with deep-learning techniques, the "Unsupervised Representation Learning" section).

Learning from unreliable or unbalanced resources

In contrast to both the no- and limited-resource techniques, which address the speech data quantity challenge, this section focuses on the methods that aim to tackle the speech data quality challenge. In particular, it covers techniques designed to operate in the presence of unreliable or unbalanced resources.

Data selection

Data quantity and diversity are both vitally important properties when building a robust ASA system. However, they can introduce a range of confounding factors. For example, speech utterances that are severely distorted by noise might be present in a prototypical data set. Owing in part to a lack of annotators' concentration, these data are often improperly labeled or even mislabeled. This gives rise to the necessity of data selection to discard such garbage data, as accurate decisions made by a pattern recognition engine are largely related to high-quality training data.

The goal of data selection is to select a smaller data source S that is most representative (i.e., most informative) of the entire data L, i.e., $S = DS(L)$ and $S \subseteq L$, thus omitting any superfluous or garbage data. The concept of data selection discussed in this section differs from that for AL or SSL, which is carried out on unlabeled data (see the "Semisupervised Learning" and "Active Learning" sections). It also differs from feature selection methods (e.g., filter or wrapper selection), which select the most informative features for a particular ASA task. Instead, the data selection techniques reviewed are designed to select labeled samples or instances that will serve as learning units.

Within the ASA literature, Wu et al. [100] selected the samples that had a uniform distribution across speech units (i.e., words and phonemes) by the principle of maximum entropy for ASR. The experimental results presented indicate that a system trained on a 150-h selection of data could achieve competitive results with a system trained on the full 840-h data set.

When performing subjective ASA recognition tasks (e.g., emotion recognition), a learning and testing target has to be generated usually by fusing the labels of multiple annotators to reduce subjectivity. In addressing the unreliable label problem, Erdem et al. [101] performed the RANSAC data selection algorithm to remove potentially mislabeled instances when training a model, and obtained better emotion recognition performance. This algorithm operates in an iterative fashion. First, it uses a small subset of the data to determine the initial model parameters. Then, the unused data instances are tested against this model, and those that fit the model within a predefined tolerance, denoted as ϵ, are considered to be a part of the consensus set. When the size reaches a predefined limit, the model parameters are updated using all of the consensus data and initial data. This procedure is repeated several times. More recently, Zhang et al. [102] reported that annotation reliability can be assessed using the human-agreement level among multiple annotators. Data with a low human-agreement level are considered to be mislabeled data and are removed from the data set.

TL approaches can be mostly grouped into one of three categories according to the properties of the knowledge transferred: instance-, feature-, and model-based TL.

Data balancing

When collecting data for a specific ASA task, such as modeling speaker states (e.g., affection or intoxication) or characteristics (e.g., likeability), one often faces issues relating to class scarcity. While interesting speech samples are required, the majority of the ubiquitous speech data are essentially neutral. This can result in highly imbalanced class distributions and recognition systems that perform poorly when attempting to recognize the target classes [103].

Numerous studies in the context of machine learning have tackled this issue by data balancing [103], with the purpose of balancing the data distribution over classes, i.e., $L_{\text{bl}} = L_1 \cup L_2 \ldots \cup L_n$ where $L_1, L_2, \ldots, L_n$ denote labeled data from n different classes that contain approximately the same amount of data. Among the methods proposed, data sampling is seen as a simple and efficient method. Data sampling is the process of either repeating preexisting data, regenerating new data to modify the imbalanced data distribution, or randomly removing part of the data to produce a data set with a more balanced class distribution.

One common method is random sampling, either by oversampling (i.e., upsampling) or by undersampling (i.e., downsampling). The former approach essentially involves randomly selecting a subset of instances L'_{min} in the minority class L_{min} and adding them back into the original training set L, $L = L \cup L'_{\text{min}}$. In contrast, the latter technique involves the random selection of a subset of instances L'_{maj} in the majority class L_{maj} and removing them from the original training set L, $L = L \setminus L'_{\text{maj}}$. However, this process may result in a loss of important information pertaining to the majority class.

Another frequently used and effective method for data sampling is SMOTE [104]. The underlying idea is the creation of a new set of artificial examples belonging to the minority class. Data sampling has been widely used for computational paralinguistics with notable effects [17], [105]. Even in ASR systems, balancing the sample distributions among all phonemes has been shown to outperform the baseline by a large margin [106].

Learning from unmatched resources

Conventional machine-learning approaches operate under the assumption that instances from both the source and the target domains are independent and identically distributed. However, in real-world scenarios, this is very rarely the case; one will inevitably encounter the problem of distribution mismatch (also known as the *data set bias*) or covariate shift between the data in the target and source domains (i.e., $\mathcal{S} \neq \mathcal{T}$). Such discrepancies often give rise to a substantial downgrade in the performance of affected speech analysis systems. TL is a potential solution to bridging the mismatch gap.

The objective of TL is to improve the predictive function in the target domain $\mathcal{T}$ using the knowledge from a different but related source domain $\mathcal{S}$ (Figure 4). A wide range of TL approaches have been proposed in the machine-learning and data-mining literature. TL has also been applied to many ASA tasks, including low-resource language ASR, speaker adaptation, and emotion recognition.

TL approaches can be mostly grouped into one of three categories according to the properties of the knowledge transferred: instance-, feature-, and model-based TL. These

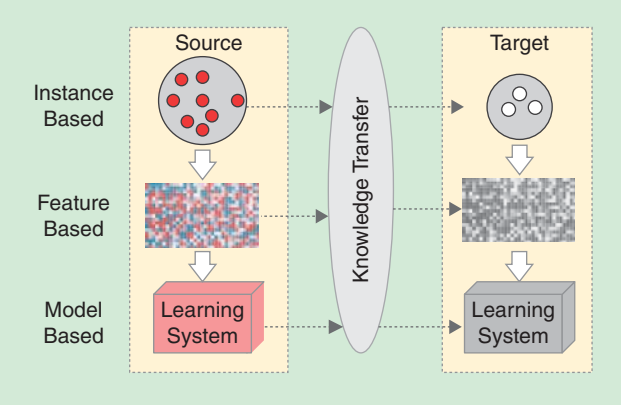


FIGURE 4. An illustration of TL: knowledge learned in the source domain is used to aid analysis in the target domain. This transfer can take place at either the instance, feature, or model level.

Table 2. Selected TL studies on the unmatched speech resource.

Publications	Types	Approaches	Models	Applications	Databases and Languages
Hassan et al. 2013 [111]	Instance	KMM, KLIEP, uLSIF	SVM	ER	FAU AEC (Ge)
Doulaty et al. 2015 [113]	Instance	Submodular data selection	DNN	ASR	Data collected in six settings
Narayanan and Wang 2013 [115]	Feature	Denoising	DNN	ASR	Aurora-4
Deng et al. 2013 [116]	Feature	SAE	SVM	ER	Six emotional corpora
Kocscor and Tóth 2004 [117]	Feature	KPCA, KLDA	GMM, ANN, etc.	Vowels/phoneme classification	Hungarian (Hu), TIMIT (En)
Jafari and Plumbley 2011 [118]	Feature	Sparse coding	/	Speech representation/ denoising	Freesound
Dahl et al. 2012 [51]	Feature	DNN, signal task	DNN–HMM	ASR	Bing mobile voice (En)
Amodei et al. 2015 [7]	Feature	CNN, signal task	CTC–RNN	ASR	English (En) and Mandarin (Ma)
Heigold et al. 2013 [119]	Feature	SHL–DNN, multitask	Softmax layer	Multi-/cross lingual ASR	Data in various languages
Huang et al. 2013 [120]	Feature	SHL–DNN, multitask	Softmax layer	Multi-/cross lingual ASR	English (En) and Mandarin (Ma)
Miao et al. 2015 [121]	Feature	SAT–DNN, *i*-vector	DNN	ASR	TEDLIUM (En)
Deng et al. 2014 [122]	Feature	SHL–DNN	SVM	ER	Three emotional corpora
Giri et al. 2015 [123]	Feature	SHL–DNN	DNN	Robust ASR	REVERB Challenge corpus (En)
Leggetter and Woodland 1995 [124]	Model	MLLR	GMM–HMM	ASR	ARPA RM (En)
Deng et al. 2014 [112]	Model	DAE, multitask	SVM	ER	Three emotional corpora

En/Ge/Hu/Ma: English/German/Hungarian/Mandarin; ER: emotion recognition; uISIF: unconstrined least-squares importance fiting; KLDA: kernal linear discriminant analysis; SHL: shared hidden layer; SAT: speaker adaptation training; DAE: denoising autoencoder; KPCA: kernel principal components analysis.

approaches as well as data agglomeration are elaborated upon in the following. These sections are intended to be a succinct overview of these techniques for ASA. For a more general survey of TL, see [20] and [21]. A selection of typical TL studies for ASA are listed in Table 2.

Instance-based TL

Instance-based TL assumes that certain subsets of the data in the source domain can be used for learning in the target domain by means of reweighting. Instance-based TL essentially assigns more weight to those source domain data that are similar in terms of distribution to the target data, and less weight to those that poorly reflect the distribution of the target data. The technique of weighting the input data based on the target data is known as *importance weighting* for covariate shift or sample selection bias. With the aim of minimizing the expected classification error, the estimation of the importance weights β is achieved as a ratio calculation problem:

$$\beta(\mathbf{x}) = \frac{P_{\mathcal{S}}(\mathbf{x})}{P_{\mathcal{T}}(\mathbf{x})}, \tag{13}$$

where $P_{\mathcal{S}}(\mathbf{x})$ and $P_{\mathcal{T}}(\mathbf{x})$ are the probability densities of the source and target domain data, respectively [107].

Instance-based TL assumes that certain subsets of the data in the source domain can be used for learning in the target domain by means of reweighting.

The most straightforward approach to calculating this density ratio is to directly estimate the target and source densities separately. However, this approach tends to perform poorly because of the inherent difficulty of density estimation, particularly in high-dimensional cases. In this regard, instance-based TL techniques, which estimate the importance ratio without estimating the densities, have been proposed. For example, Huang et al. [108] proposed a kernel-based method known as *kernel mean matching* (*KMM*). It reweights the instances by matching the means between the source domain data and the target domain data in a reproducing-kernel Hilbert space. The downside of KMM is that its performance is highly dependent on the choice of hyperparameters (model selection), which need to be heuristically tuned.

To overcome this issue, Sugiyama et al. [109] introduced the Kullback–Leibler importance estimation procedure (KLIEP) algorithm. KLIEP estimates the importance ratio by minimizing the Kullback–Leibler divergence between the original target data density and its corresponding estimation. Owing to the convex property of the involved optimization problem, the KLIEP algorithm can obtain unique global solutions. In addition, the tuning parameters can be objectively optimized, based on a variant of cross validation. While KLIEP is seemingly

more advantageous than KMM, it is actually less computationally efficient because of the high linearity of the objective functions to be optimized.

This issue was addressed by Kanamori et al. [110] by means of least-squares importance fitting (LSIF). The LSIF algorithm formulates the direct importance estimation problem as a least-square function fitting problem: casting the optimization problem as a convex quadratic program that can be efficiently solved using a standard quadratic program solver. This algorithm was further extended to be unconstrained LSIF (uLSIF), which greatly improved the computational efficiency [110]. For emotion recognition, the approaches of KMM, KLIEP, and uLSIF have shown great success in alleviating the discrepancy between different speech resources [111], [112].

An alternative to the aforementioned approaches is binary reweighting. It selects the data from the source domain based on the data distribution to reduce the discrepancy between the source domain and the target domain. This strategy is related to the data selection strategy used for AL (see the "Active Learning" section), which can be viewed as a specific data selection case in a source-data unlabeled setting. It is also related to the data selection strategy discussed in the "Data Selection" section, which attempts to improve the quality of the data only in the target domain.

A prominent binary reweighting approach is based on using submodular functions to simulate the acoustic similarity between the target and source domain data.

A prominent binary reweighting approach is based on using submodular functions to simulate the acoustic similarity between the target and source domain data [113], [114]. The process identifies a subset L' of the complete source data set $L_{\mathcal{S}}$, so that any subsequent subset L'' added to this selected subset will not increase the value of the submodular function f, i.e., $L' = \arg\max\{f(L' \cup L'') < f(L')$, where $L' \subseteq L, L'' \subseteq L \setminus L'\}$. In doing this, only the positive transfer is exploited across domains. In ASA, submodular function-based data selection has been extensively evaluated for multidomain speech recognition and has shown superior performance [113], [114].

Feature-based transfer learning

The goal of feature-based TL approaches is to find a transformation function $\Phi(\cdot)$ that can be used to convert the source feature space and/or target feature space into an approximately matched distribution space while preserving the important properties of the original data. Mathematically, this can be expressed as

$$P(\Phi_{\mathcal{T}}(X_{\mathcal{T}})) \approx P(\Phi_{\mathcal{S}}(X_{\mathcal{S}})), \tag{14}$$

or

$$P(Y_{\mathcal{T}} | \Phi_{\mathcal{T}}(X_{\mathcal{T}})) \approx P(Y_{\mathcal{S}} | \Phi_{\mathcal{S}}(X_{\mathcal{S}})). \tag{15}$$

In achieving this, two possible strategies exist: asymmetric and symmetric strategies. The asymmetric strategy keeps either the source or target feature space unchanged, and maps the other one onto it (i.e., $\Phi_{\mathcal{T}}: \mathcal{T} \rightarrow \mathcal{S}$ or $\Phi_{\mathcal{S}}: \mathcal{S} \rightarrow \mathcal{T}$). By contrast, the symmetric strategy transforms both source and target feature spaces into a new latent one (i.e., $\Phi_{\mathcal{T}}: \mathcal{T} \rightarrow \mathcal{Z}$ and $\Phi_{\mathcal{S}}: \mathcal{S} \rightarrow \mathcal{Z}$), in which they share the same distribution and knowledge relationship.

In achieving this, two possible strategies exist: asymmetric and symmetric strategies. The process of denoising distorted (noisy) speech can make the feature space (target) of noisy speech closer to that of clean speech (source). In doing this, the cleaned speech can be evaluated by preexisting acoustic models, which are often trained on the clean speech. An emerging research trend in the speech enhancement community is to use DNNs (e.g., deep LSTM–RNNs) to map noisy speech into its clean counterpart or ratio mask on a frame-by-frame basis. Preliminary results have proved that this method is quite effective, particularly for alleviating nonstationary noise [115]. For more details of speech denoising technologies, see [125].

Apart from speech denoising, a more general TL method to reduce the database bias was proposed in [116] and is based on an SAE—an autoencoder with sparsity enforced in the hidden layer (see the "Unsupervised Representation Learning" section). This method is a fully supervised approach. First, using the target data, class-specific SAEs are trained, and then treated as the transforming models ($\Phi(\cdot)$). The source data are then fed into SAEs corresponding to its class, and thus a new source representation is constructed. In doing this, the distribution of the new source feature space is expected to be inclined to the target one. Finally, the new source data are used to train a standard classifier.

As for the symmetric strategy, early studies were mainly conducted using principal component analysis (PCA), linear discriminant analysis (LDA), and sparse coding. The goal of these approaches is to learn a low-dimensional latent feature space or a shared space. The resulting feature space can serve as a bridge for transferring meaningful knowledge from the source domain to the target domain [20]. PCA is typically used to project the data along the direction of maximal variance in an unsupervised way. LDA, or Fisher's LDA (FDA), on the other hand, is used to project the data onto a line that can maximize the distance between the means of the two classes (in a binary classification case) while minimizing the variance within each class.

Both PCA and LDA are linear transformations that limit their applicability to most real-world data. In this regard, kernel functions (e.g., Gaussian, Cauchy, and polynomial kernels) can be used in conjunction with PCA and FDA, resulting in kernal PCA (KPCA) and kernel FDA (KFDA) paradigms that transform data in a nonlinear manner. Owing to their simplicity and effectiveness, KPCA and KFDA have been widely used in the speech processing community [117]. Similarly, kernel canonical correlation analysis has been applied to cross lingual emotion recognition [126].

Sparse coding, also termed *dictionary learning*, attempts to find succinct representations (i.e., atoms or elements of the dictionary) of the input data such that the input data can be represented as a linear combination of these sparse representations [127]. Compared to the aforementioned feature transformation methods, sparse coding has been demonstrated to be able to produce a more robust signal representation in speech reconstruction and denoising tasks [118].

Conventional feature transformation approaches are typically executed at a shallow level. Recently, deep-learning approaches for feature-based TL have begun to attract a lot of of research attention. Deep learning is regarded as a natural TL paradigm; it provides a powerful capability of learning high-level abstracts or representations that are more robust against the variation of conventional speech features (i.e., log Mel-filter banks and MFCCs) over different domains [50] (see the "Unsupervised Representation Learning" section). These representative features can then be used as normal features to train conventional discriminative or generative models, such as NNs, HMMs, and SVMs. Thanks to the invariant property of these representations, they can potentially deliver remarkable performance improvements for almost all ASA tasks [7], [50], [51], [58].

Researchers have started to investigate the learning of robust representations over multiple modalities (e.g., audio and video).

In addition to the basic representation learning approaches mentioned previously, more advanced topologies have begun to emerge, which explicitly involve several related tasks in a multitask learning paradigm. Multitask learning is the process of learning multiple tasks at the same time to learn a shared representation among different tasks. Mathematically, when training the model with multiple tasks, we aim to minimize the objective function as follows:

$$\mathcal{J}(\theta_0) = \sum_{k=1}^{K} \sum_{i} L(\mathbf{x}_{ki}, y_{ki}; \theta_k) + \frac{\lambda}{2} \| \theta_0 \|^2, \tag{16}$$

where K is the number of tasks, $L(\cdot)$ denotes the loss function, and θ_0 stands for the general model parameters.

When performing deep multitask learning for multilingual or cross lingual speech recognition, it is typical to share the hidden layers across all languages [119], [120]. If learned appropriately, the hidden layers serve as increasingly complex feature transformations, sharing common hidden factors across the acoustic data from different languages. The final softmax layers, however, are not shared. Instead, each language has its own softmax layer to estimate the posterior probabilities specific to that language, using the most abstract representation from the topmost hidden layer. The strong result gained using this topology [119], [120] indicates its potential; it opens up the possibility for quickly building a high-performance recognition system for a new language using an existing multilingual DNN.

Many other deep multitask learning derivatives have been investigated to overcome the feature variation problems caused by factors such as different speaker characteristics, noisy environments, and poor recording channels. For example, Deng et al. [122] treated different corpora as different tasks for emotion recognition; Giris et al. [123] regarded noise type as an auxiliary task for speech recognition; and Seltzer and Droppo [128] treated phone label, phone text, and state context as different tasks when performing phoneme recognition. Recently, a universum autoencoder was proposed [129]. This technique uses a small amount of labeled data from the target domain and unlabeled data from a source domain to jointly minimize the reconstruction error and the universum leaning loss. Motivated by these achievements of learning representations among multiple related tasks, researchers have started to investigate the learning of robust representations over multiple modalities (e.g., audio and video) [130]. This topic, however, is beyond the scope of this overview.

Model-based transfer learning

Model-based TL, also known as *parameter-based TL*, aims to learn a new model from an existing model that has been well trained on rich source data. Unlike feature-based TL approaches, which usually transform the feature spaces, model-based TL approaches modify the pretrained model parameters (θ) to account for the differences that may exist between the domains. This can be formulated as

$$P(X_S, Y_S; \theta_S) \rightarrow P(X_T, Y_T; \theta_T) \tag{17}$$

for a generative model or

$$P(Y_S | X_S; \theta_S) \rightarrow P(Y_T | X_T; \theta_T) \tag{18}$$

for a discriminative model.

Early-stage model-based TL approaches in the speech community included maximum a posteriori (MAP) estimation and maximum likelihood linear regression (MLLR), which are designed for generative models (e.g., GMM–HMM). These techniques have been applied successively to speaker adaptation [131], where the speech from each specific speaker is supposed to be in a different domain with the initial training data. They have also been shown to be useful in computational paralinguistics tasks, such as depression detection [132].

Specifically, MAP uses the speaker-independent models (i.e., universal background models) as a prior probability distribution over the model parameters, and then performs maximum likelihood estimates by considering the model parameters obtained on the speaker-dependent data. Alternatively, MLLR calculates a set of linear regression transformations to shift both the means and the covariances in an initial Gaussian mixture HMM system so that each state in the system is more likely to have generated the speaker data the model is being adapted to [131]. Compared with MAP, MLLR requires fewer adaptive data. Aside from speaker adaptation, these methods have been applied to

other acoustic variation adoption scenarios, such as noise adaptation [125].

Due largely to the recent advancements in deep learning, discriminative model-based TL has recently become an active research topic. In deep learning, the simplest way to adjust the pretrained model parameters when adapting to a specific task is through fine-tuning. As discussed in the "Unsupervised Representation Learning" section, pretraining is a down–up unsupervised algorithm, which can be considered as a model initialization process that attempts to produce a model that has a global optimization attribute. By contrast, fine-tuning is an up–down supervised algorithm to optimize all of the NN weights jointly with the labeled target data. This procedure is usually performed using backpropagation of error derivatives [63].

In contrast to the more sophisticated TL approaches discussed, a simpler solution to utilize multiple sources of data is data agglomeration.

Another paradigm to adapt the model to the target data, the adaptive denoising autoencoder, is highly related to multitask learning [112], [133]. This paradigm is usually undertaken in two steps. In the first step, a source model is trained on the source data. In the second step, the trained model parameters are used as prior information to regularize the adaptation process of the model on the target data, so as to minimize the objective function as follows:

$$\mathcal{J}(\theta_{\mathcal{T}}) = \sum_{i=1}^{n_{\mathcal{T}}} L(\mathbf{x}_i, y_i; \theta_{\mathcal{T}}) + \frac{\lambda}{2} \| \theta_{\mathcal{T}} - \beta \theta_{\mathcal{S}} \|^2, \tag{19}$$

where $n_{\mathcal{T}}$ is the number of labeled target data, $L(\cdot)$ denotes the loss function on the target data, $\theta_{\mathcal{S}}$ represents the well-trained model on the source data (source model), $\theta_{\mathcal{T}}$ denotes the expected new model on the target data (target model), and β is the adaptation coefficient. Since the discrepancy between the source and target models is explicitly considered as a penalty term in the objective function, this approach is also known as *regularized adaptation* [133]. In emotion recognition applications, this approach has started to show promising results [112]. Note that such model-based multitask learning paradigms differ from the feature-based approaches covered in the "Learning from Unmatched Resources" section, where the model is trained in only one step by calculating the joint loss of all of the tasks in the objective function [see (16)].

Data agglomeration

In contrast to the more sophisticated TL approaches discussed, a simpler solution to utilize multiple sources of data is data agglomeration [134]. In this approach, one or more source databases are directly concatenated with the target database to form a large-size data pool $P = L_{\mathcal{T}} \cup L_{\mathcal{S}_1} \cup \ldots \cup L_{\mathcal{S}_k}$. This approach is suitable only when the various data sources are for similar tasks and share a common feature set.

To help ease any potential database biases, it is desirable to apply 1) normalization techniques such that the scattered feature spaces can be unified into a shared one and 2) task mapping to retain label consistency. The three normalization methods frequently applied in the literature are centering, min–max normalization, and standardization. Applied not only to each corpus separately (i.e., before data agglomeration), these methods can be also used after building a joint training set from multiple databases. Thanks to these normalization approaches, data agglomeration has been frequently applied to, e.g., emotion recognition [134]. As for task mapping, it is necessary to find the relationship between different tasks. For example, in emotion recognition, the prototypical emotions (e.g., anger, contempt, disgust, fear, interest, joy, sadness, and surprise) can be mapped onto the emotional dimensions of arousal and valence [134].

Conclusions and challenges for future work

To continue building on the success of machine-learning methods for ASA, there is a need for large amounts of labeled data. However, the work of collecting such data is costly and time consuming. Clever engineering can go a long way toward solving this problem by helping to leverage unlabeled, unreliable, or unmatched data. Motivated by this, we systematically presented an overview of the very recent and prominent techniques that intend to semiautonomously enrich the data quantity and enhance the data quality.

Crowdsourcing was discussed as an efficient data annotation approach, with the caveat that it requires quality control management. The integration of crowdsourcing with AL or CL strategies to intelligently and dynamically select data for labeling has the potential to further reduce the annotation workload and improve overall data quality.

Spoken-term detection and discovery and related means of retrieval of speech-related phenomena were discussed in relation to addressing the sparse data challenge. While these techniques can automatically find patterns in speech utterances without any labeled resource, the associated computational complexity limits their application to smaller databases. Reducing the computing complexity of these techniques is an essential direction of future research. Other techniques discussed on the sparse data challenge were data augmentation and speech synthesis. These techniques can artificially generate labeled speech data in a limited-labeled-resource setting. A key concern about their ongoing use is how to guarantee that the speech samples generated have a positive effect on the analysis being performed. Research into identifying task-invariant features has been identified as one potential solution in this regard.

With its capability to leverage information from large-scale unlabeled data, deep URL has delivered breakthrough results in a variety of ASA tasks. Future research efforts, particularly those focused on network construction strategies, are expected to increase the generalizability of the extracted features and thus improve on the already impressive capabilities of this paradigm. AL, SSL, and CL are other efficient techniques to take advantage of unlabeled data. In this regard, we identified the

integration of SSL and deep learning as a particularly promising future research direction.

To handle the unreliable-data challenge, data selection and data-balancing techniques were also reviewed. Despite the conventionality of the reviewed algorithms, dynamically selecting and balancing data is of great importance to the machine-learning process. The role and importance of these well-practiced techniques in relation to deep learning are still being established.

To deal with the unmatched data challenge, TL strategies and data agglomeration were discussed. TL in particular, owing to its effectiveness, has attracted increasing amounts of research attention. However, when improperly used, these techniques substantially degraded overall system performance. Therefore, how to achieve positive transfer while preventing negative transfer between appropriately related tasks is an important and open research issue.

Although great opportunities are offered by the techniques reviewed, many additional risks may be brought to light through their practical application. For example, with the growing popularity of the use of microphones, the Internet, crowdsourcing, and cloud computing, personal speech signals easily run the risk of being disclosed to the public domain. Furthermore, from such data it is largely possible to extract confidential speaker information, such as a speaker's age, gender, or identity. Therefore, how to best protect the security and privacy of users has become a major area of concern in this field [135].

A potential solution in this regard is a distributed recognition system, such as the one proposed for computational paralinguistics in [136]. In this system, functionals are applied over the LLDs to extract features. These statistical features, rather than the LLDs or the raw signals, are transmitted from the client side to the server side. The procedure of generating these feature vectors is irreversible. Therefore, as the LLDs cannot be reconstructed, the contents of the original speech signals are protected. Recently, a decentralized SSL paradigm was proposed in [137], in which privacy-preserving matrix completion algorithms are used, so that only learned knowledge is transferred between different clients, while the raw data are incommutable. However, as these approaches cannot fully guarantee client security and privacy or maintain the original performance, continued research addressing privacy concerns is required.

The techniques discussed in this article are mainly applied in an offline manner. However, the realistic application of a specific task offers the opportunity to collect truly massive amounts of real-world data in an online fashion. For example, Google reported that 55% of teenagers and 41% of adults in the United States [138] used their voice search more than once a day in 2014. Hence, research is needed into techniques to dynamically make use of future data to enhance the adaptiveness of preexisting models to various speakers, environments, and tasks. Such techniques are commonly referred to as *online* and *incremental* learning [139], [140].

Finally, the recent developments in dialog management systems, the computerized spoken language understanding and generation of natural and meaningful responses during speech-based human–computer interactions, means it is now more feasible than ever to explore cues extracted from an entire conversation process to aid ASA systems. Such cues could indicate the correctness of previously performed analyses and as such would be considered a form of reward or punishment information. This information could be sequentially exploited using reinforcement learning strategies to dynamically update the decision mechanism of the predictive model. Deep reinforcement learning, in particular, has become an active and growing research topic in machine learning [141]. But despite being widely applied in related fields, such as dialog management, research into reinforcement learning for ASA is currently in its infancy. We firmly believe that research into deep reinforcement learning has the potential to move ASA technologies out of controlled laboratory settings and into diverse, practical everyday environments leading to more intelligent (even emotionally and socially intelligent) and adaptive ASA systems.

Despite these risks and challenges, the techniques reviewed in this article will play a key role in opening up new research opportunities to explore the value of big unlabeled, unreliable, and unmatched speech data. It is our strong belief that the continued growth in the research and applications discussed will facilitate the emergence of novel techniques to fill the gap between no-labeled-resource and reliable big data and usher in the next generation of ASA technologies.

Acknowledgments

This work was supported by the European Union's Seventh Framework Program through ERC Starting Grant 338164 (iHEARu), and by the Horizon 2020 Program through Research Innovation Action 688835 (DE-ENIGMA).

Authors

Zixing Zhang (zixing.zhang@uni-passau.de) received his B.S. degree from the Chinese Agricultural University in 2007, his M.S. degree from the Beijing University of Posts and Telecommunications, China, in 2010, and his Ph.D. degree from the Technische Universität München, Germany, in 2015. Currently, he is a postdoctoral researcher at the University of Passau, Germany. His research interests lie mainly in deep, semisupervised, active, and multitask learning; in the applications of human state and trait analysis from speech; and in robust automatic speech recognition. He is a Member of the IEEE.

Nicholas Cummins (nicholas.cummins@uni-passau.de) received his B.Eng. degree (first-class honors) in electrical engineering from the University of New South Wales (UNSW), Sydney, Australia and his Ph.D. degree in electrical engineering from UNSW in February 2016. His Ph.D. dissertation investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He is currently a postdoctoral researcher at the Chair of Complex and Intelligent Systems, University of Passau, Germany. His research interests include affective and behavioral computing. He is a Member of the IEEE.

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

Björn Schuller (schuller@ieee.org) received his diploma, doctoral degree, and habilitation degree in electrical engineering and information technology from the Technische Universität München, Germany in 1999, 2006, and 2012, respectively. He is a reader in machine learning in the Department of Computing at Imperial College, London, United Kingdom, and a full professor and head of the Chair of Complex and Intelligent Systems, University of Passau, Germany, where he previously headed the Chair of Sensor Systems. He is a Senior Member of the IEEE.

References

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

[2] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, and J. H. L. Hansen, "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 49–60, Nov. 2016.

[3] B. W. Schuller, "The computational paralinguistics challenge," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 97–101, July 2012.

[4] C. Moseley, *Atlas of the World's Languages in Danger*, 3rd ed. Paris: Unesco Publishing, 2010.

[5] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[6] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Process.*, vol. 7, no. 3–4, pp. 197–387, June 2014.

[7] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. Int. Conf. Machine Learning (ICML)*, New York, 2016, pp. 173–182.

[8] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[10] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[11] M. Harper. IARPA Babel program. Intelligence advanced research projects activity, Office of the Director of National Intelligence, Washington, D.C. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[12] B. W. Schuller, "Speech analysis in the big data era," in *Text, Speech, and Dialogue* (Lecture Notes in Computer Science, vol. 9302), P. Král and V. Matoušek, Eds. Berlin: Springer-Verlag, 2015, pp. 3–11.

[13] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3169–3173.

[14] M. R. Robertson. (2015, Nov. 13). 500 hours of video uploaded to YouTube every minute. *Tubular Insights*. [Online]. Available: http://www.reelseo.com/hours-minute-uploaded-youtube

[15] M. Eskénazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Hoboken, NJ: Wiley, 2013.

[16] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon, "Crowdsourcing for difficult transcription of speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, 2011, pp. 535–540.

[17] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[18] Z. Zhang, *Semi-Autonomous Data Enrichment and Optimisation for Intelligent Speech Analysis*. Munich, Germany: Verlag Dr. Hut, 2015.

[19] A. Nagórski, L. Boves, and H. J. Steeneken, "Optimal selection of speech data for automatic speech recognition systems," in *Proc. INTERSPEECH*, Denver, CO, 2002, pp. 2473–2476.

[20] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proc. Asia-Pacific Signal and Information Processing Assoc. Annu. Summit and Conf. (APSIPA)*, Hong Kong, China, 2015, pp. 1225–1237.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[22] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[23] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP)*, Honolulu, HI, 2008, pp. 254–263.

[24] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Proc. Human Language Technologies: 2010 Annu. Conf. North American Chapter Assoc. Computational Linguistics*, Los Angeles, 2010, pp. 207–215.

[25] S. Hantke, T. Appel, F. Eyben, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 891–897.

[26] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, et al. "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8111–8115.

[27] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Berkeley, CA, 2010, pp. 312–317.

[28] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *Proc. W3C workshop on Emotion Markup Language (EmotionML)*, Paris, 2010, pp. 1–5.

[29] J. Ledlie, B. Odero, E. Minkov, I. Kiss, and J. Polifroni, "Crowd translator: On building localized speech recognizers through micropayments," *ACM SIGOPS Operating Syst. Rev.*, vol. 43, no. 4, pp. 84–89, Jan. 2010.

[30] C-Y. Lee and J. R. Glass, "A transcription task for crowdsourcing with automatic quality control," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3041–3044.

[31] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[32] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 410–415.

[33] H. Wang, T. Lee, C. C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8545–8549.

[34] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8515–8519.

[35] X. Anguera, "Method and system for improved pattern matching," EP Patent EP12 382 508, 2012.

[36] Y. Chung, C. Wu, C. Shen, H. Lee, and L. Lee, "Audio Word2Vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 765–769.

[37] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, Apr. 2016.

[38] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 4366–4369.

[39] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 5532–5536.

[40] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, Atlanta, GA, 2013.

[41] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 23, no. 9, pp. 1469–1477, Sept. 2015.

[42] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 1420–1424.

[43] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3586–3589.

[44] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3214–3218.

[45] B. Milde and C. Biemann, "Using representation learning and out-of-domain data for a paralinguistic speech task," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 904–908.

[46] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *Int. J. Speech Technol.*, vol. 15, no. 3, pp. 313–323, June 2012.

[47] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust ASR," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, Merano, Italy, 2009, pp. 205–210.

[48] Z. Ling, S. Kang, H. Zen, A. Senior, M. Schuster, X. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[49] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jan. 2006.

[50] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[51] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[52] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learning Res.*, vol. 11, pp. 625–660, Mar. 2010.

[53] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Machine Learning (ICML)*, Helsinki, Finland, 2008, pp. 1096–1103.

[54] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A-r. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1692–1695.

[55] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten ZIP code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[56] O. Abdel-Hamid, A-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[57] D. Hau and K. Chen, "Exploring hierarchical speech representations with a deep convolutional neural network," in *Proc. 11th U.K. Workshop on Computational Intelligence (UKCI)*, Manchester, U.K., 2011, pp. 37–42.

[58] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Advances Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009, pp. 1096–1104.

[59] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Int. Conf. Machine Learning (ICML)*, New York, 2009, pp. 609–616.

[60] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[61] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Machine Learning (ICML)*, Lille, France, 2015, pp. 843–852.

[62] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Workshop Syntax, Semantics and Structure Statistical Translation (SSST)*, Doha, Qatar, 2014, pp. 103–111.

[63] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[64] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 437–440.

[65] A-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[66] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1695–1699.

[67] Y. Liu, T. Fu, Y. Fan, Y. Qian, and K. Yu, "Speaker verification with deep features," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, Beijing, 2014, pp. 747–753.

[68] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5688–5691.

[69] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Deep learning for emotional speech recognition," in *Pattern Recognition*, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. A. Olvera-Lopez, J. Salas-Rodríguez, and C. Y. Suen, Eds. MCPR 2014. *Lecture Notes in Computer Science*, vol. 8495. Berlin: Springer-Verlag, 2014, pp. 311–320.

[70] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2013, pp. 3687–3691.

[71] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2725–2728.

[72] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 23–31, Jan. 2005.

[73] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, "Speaker identification using semi-supervised learning," in *Proc. 17th Int. Conf. Speech and Computer (SPECOM)*, Athens, Greece, 2015, pp. 389–396.

[74] R.-C. Hsiao, T. Ng, F. Grézl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 440–445.

[75] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6704–6708.

[76] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8505–8509.

[77] X. Cui, J. Huang, and J.-T. Chien, "Multi-view and multi-objective semi-supervised learning for HMM-based automatic speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1923–1935, Sept. 2012.

[78] Y. Liu and K. Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1946–1956, Nov. 2016.

[79] G. Riccardi and D. Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 504–511, July 2005.

[80] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Taibei, China, 2009, pp. 4721–4724.

[81] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, A. Laurent, V.-B. Le, and A. Messaoudi, "Active learning based data selection for limited resource STT and KWS," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 47–53.

[82] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, 2010, pp. 4350–4353.

[83] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. INTERSPEECH*, Portland, OR, 2012, pp. 362–365.

[84] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller, "On rater reliability and agreement based dynamic active learning," in *Proc. 6th Biannu. Conf. Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 70–76.

[85] G. Riccardi and D. Z. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proc. INTERSPEECH*, Geneva, Switzerland, 2003, pp. 1825–1828.

[86] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Commun.*, vol. 52, no. 7, pp. 652–663, Aug. 2010.

[87] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Comput. Speech & Language*, vol. 24, no. 3, pp. 433–444, July 2010.

[88] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin, Madison, Tech. Rep. TR 1530, 2006.

[89] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2000.

[90] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, 2011, pp. 523–528.

[91] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5185–5189.

[92] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Computational Learning Theory (COLT)*, Madison, WI, 1998, pp. 92–100.

[93] H. Xu, H. Su, C. Ni, X. Xiao, H. Huang, E. S. Chng, and H. Li, "Semi-supervised and cross-lingual knowledge transfer learnings for DNN hybrid acoustic models under low-resource conditions," in *Proc. INTERSPEECH*, San Francisco, 2016, pp. 1315–1319.

[94] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin, Madison, Tech. Rep. TR 1648, 2009.

[95] N. Roy and A. McCallum, "Toward optimal active learning through Monte Carlo estimation of error reduction," in *Proc. 18th Int. Conf. Machine Learning (ICML)*, Williamstown, MA, 2001, pp. 441–448.

[96] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1323– 1331, Aug. 2010.

[97] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learning*, vol. 28, no. 2-3, pp. 133–168, Aug. 1997.

[98] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proc. Int. Conf. Machine Learning (ICML)*, Madison, WI, 1998, pp. 359–367.

[99] I. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. Int. Conf. Machine Learning (ICML)*, Sydney, Australia, 2002, pp. 435–442.

[100] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Kyoto, Japan, 2007, pp. 562–565.

[101] C. E. Erdem, E. Bozkurt, E. Erzin, and A. T. Erdem, "RANSAC-based training data selection for emotion recognition from spontaneous speech," in *Proc. 3rd Int. Workshop on Affective Interaction in Natural Environments (AFFINE)*. New York, 2010, pp. 9–14.

[102] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. 5th Int. Workshop Emotion Social Signals, Sentiment & Linked Open Data (satellite of LREC 2014)*, Reykjavik, Iceland, 2014, pp. 21–26.

[103] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

[104] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intell. Res.*, vol. 16, pp. 321–357, June 2002.

[105] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.

[106] A. I. García-Moral, R. Solera-Ureña, C. Peláez-Moreno, and F. Díaz-de María, "Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 468–481, Mar. 2011.

[107] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, Oct. 2000.

[108] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Proc. Advances Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2006, pp. 601–608.

[109] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007, pp. 1433–1440.

[110] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *J. Mach. Learning Res.*, vol. 10, pp. 1391–1445, July 2009.

[111] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[112] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sept. 2014.

[113] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2897–2901.

[114] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3311–3315.

[115] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[116] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, Geneva, Switzerland, 2013, pp. 511–516.

[117] A. Kocsor and L. Tóth, "Kernel-based feature extraction with a speech technology application," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2250–2263, Aug. 2004.

[118] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1025–1031, Sept. 2011.

[119] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8619–8623.

[120] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7304–7308.

[121] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015.

[122] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 4818–4822.

[123] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5014–5018.

[124] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[125] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[126] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5800–5804.

[127] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, June 1996.

[128] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6965–6969.

[129] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[130] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Machine Learning (ICML)*, Bellevue, WA, 2011, pp. 689–696.

[131] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001, pp. 11–19.

[132] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Commun.*, vol. 75, pp. 27–49, Dec. 2015.

[133] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. I–237–I–240.

[134] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 119–131, July 2010.

[135] S. Y. Kung, "Compressive privacy: From information/estimation theory to machine learning," *IEEE Signal Process. Mag.*, vol. 34, no. 1, pp. 94–112, Jan. 2017.

[136] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Trans. Affective Comput.*, vol. 5, no. 4, pp. 406–417, Oct. 2014.

[137] R. Fierimonte, S. Scardapane, A. Uncini, and M. Panella, "Fully decentralized semi-supervised learning via privacy-preserving matrix completion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–13, 2016.

[138] S. Huffman. (2014, Oct. 14). OMG! Mobile voice survey reveals teens love to talk. [Online]. Available: https://googleblog.blogspot.de/2014/10/omg-mobile-voice-survey-reveals-teens.html

[139] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha, and L. Zhao, "Practical speech emotion recognition based on online learning: From acted data to elicited data," *Math. Problems in Eng.*, vol. 2013, pp. 9, June 2013.

[140] W. Ainsworth and S. Pratt, "Feedback strategies for error correction in speech recognition systems," *Int. J. Man-Mach. Stud.*, vol. 36, no. 6, pp. 833–842, June 1992.

[141] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, et al. "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

SP

John H.L. Hansen,
Carlos Busso, Yang Zheng,
and Amardeep Sathyanarayana

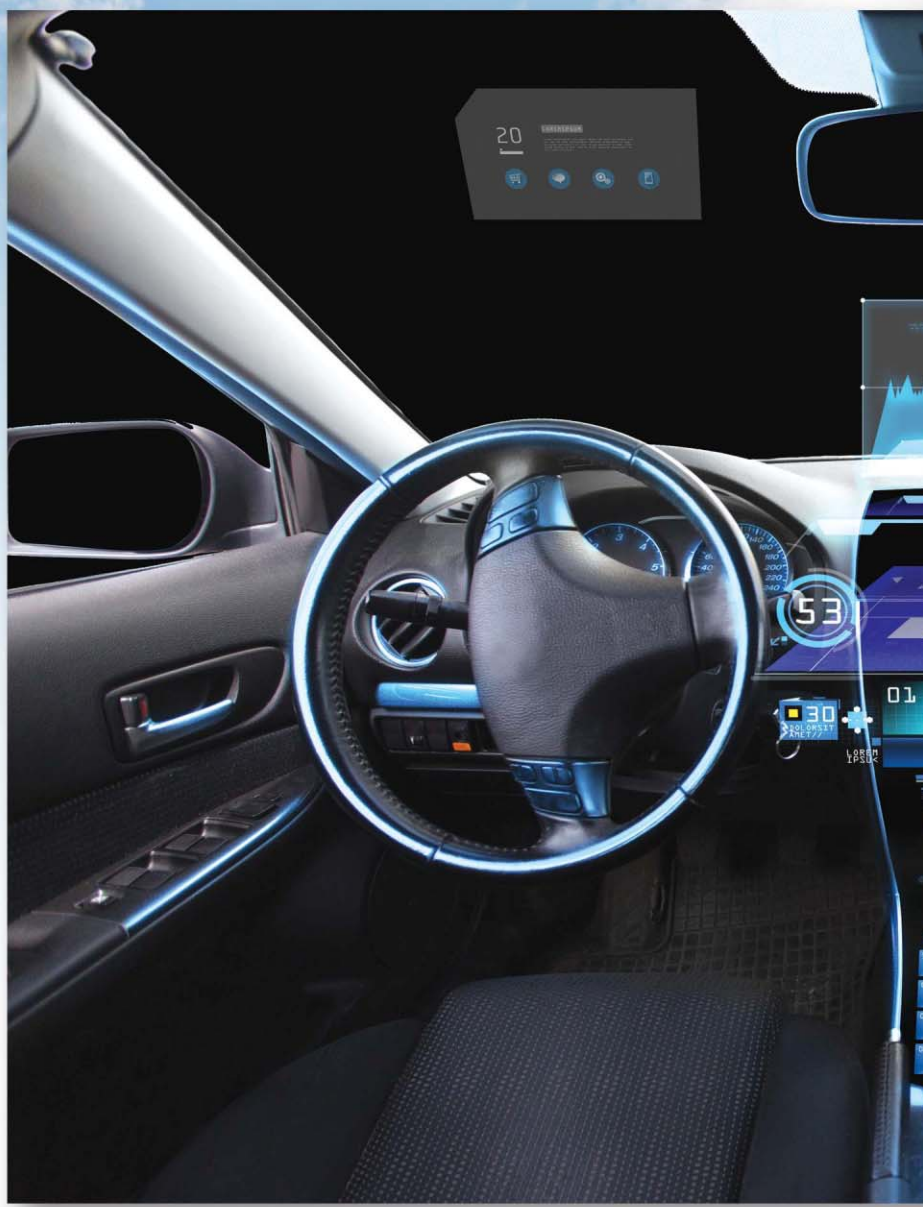


IMAGE LICENSED BY INGRAM PUBLISHING, SKY—©GRAPHIC STOCK

Vehicle technologies have advanced significantly over the past 20 years, especially with respect to novel in-vehicle systems for route navigation, information access, infotainment, and connected vehicle advancements for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) connectivity and communications. While there is great interest in migrating to fully automated, self-driving vehicles, factors such as technology performance, cost barriers, public safety, insurance issues, legal implications, and government regulations suggest it is more likely that the first step in the progression will be multifunctional vehicles. Today, embedded controllers as well as a variety of sensors and high-performance computing in present-day cars allow for a smooth transition from complete human control toward semisupervised or assisted control, then to fully automated vehicles. Next-generation vehicles will need to be

Driver Modeling for Detection and Assessment of Distraction

Examples from the UTDrive Test Bed

Digital Object Identifier 10.1109/MSP.2017.2699039
Date of publication: 11 July 2017

The need for driver modeling research

Over the past few years, there has been a significant effort in establishing smart cars that have the capability to achieve self/autonomous driving for their passengers or the passive operator. While great strides in autonomous driving will continue, greater research and understanding is needed regarding driver modeling as we transition from full-driver control to various levels of assistive-through-fully automated vehicles. The ability for smart cars to seamlessly move back and forth between completely automated, semiautomated, semiassistive, and unassisted cars remains a major challenge. In this article, we consider an overview of the recent advancements in driver modeling to assess driver status, including the detection and assessment of driver distraction when the vehicle is operated in a user-controlled scenario. The recent large-scale data collection undertaken by the United States Transportation Research Board—Strategic Highway Research Program [1], [91] will provide an enormous (+2 Petabytes) data set of naturalistic data. The ability for researchers to mine this corpus to develop better models of driver status will offer new insights into next-generation smart vehicles, which have the capability of migrating between being completely user controlled to fully autonomous.

more active in assessing driver awareness, vehicle capabilities, and traffic and environmental settings, plus how these factors come together to determine a collaborative safe and effective driver–vehicle engagement for vehicle operation. This article reviews a range of issues pertaining to driver modeling for the detection and assessment of distraction. Examples from the UTDrive project are used whenever possible, along with a comparison to existing research programs. The areas addressed include 1) understanding driver behavior and distraction, 2) maneuver recognition and distraction analysis, 3) glance behavior and visual tracking, and 4) mobile platform advancements for in-vehicle data collection and human–machine interface. This article highlights challenges in achieving effective modeling, detection, and assessment of driver distraction using both UTDrive instrumented vehicle data and naturalistic driving data.

An extensive amount of research and development is currently being conducted by many laboratories in the United States, Japan, Germany, Sweden, South Korea, and other countries; it is therefore not possible to provide exhaustive coverage of all significant advancements. Instead, the goal here is to provide a representative look at the topic of driver modeling, focusing on how advancing technologies impact driver distraction. This includes a range of signal processing technologies related to controller area network (CAN) bus analysis, image/video processing, speech/audio for human–machine interaction, and other advancements leading to current and future intelligent assistance in the vehicle.

A recent *IEEE Signal Processing Magazine* special issue, "Smart Vehicle Technologies: Signal Processing on the Move" [2], considered a range of topics for smart vehicles advancements that included driver behavior modeling using on-road driving data [3], driver status monitoring systems [4], smart driver monitoring [5], conversational in-vehicle dialog systems [6], active noise control in cars [7], and coordinated autonomous vehicles [8]. In this article, we provide several complementary highlights to these excellent overview articles. Several experiments/data sets have collected information on driver behavior analysis [9]–[11]. For the sake of illustration, the UTDrive naturalistic driving data set [12] has been conducted by the Center for Robust Speech Systems (CRSS)-UTDrive since 2006, with the interest of understanding driver behavior and distraction from multichannel sensor data (see Figure 1) [13]. Here, we focus on current advancements, past efforts, and directions for future research. Examples stemming from the UTDrive project are highlighted as examples, as well as efforts from the Virginia Tech Transportation Institute, the University of Michigan Transportation Research Institute, the University of California, San Diego, plus studies conducted in Europe, Japan, and South Korea [14].

The ability for smart cars to seamlessly move back and forth between completely automated, semiautomated, semiassistive, and unassisted cars remains a major challenge.

Understanding driver behavior and driving distraction

Driver activities performed within the vehicle can be broadly classified into primary tasks that are essential for operating and directing the course of a vehicle in a given environment and secondary tasks that are not essential or related to the primary task of driving. Secondary tasks divert drivers' primary attention of driving and degrade their driving performance. The deterioration is directly attributed to the driver (distraction, inattention), the vehicle (condition, familiarity), or the surrounding environment (traffic, weather). Both driver distraction and driver inattention are frequently occurring events in a car.

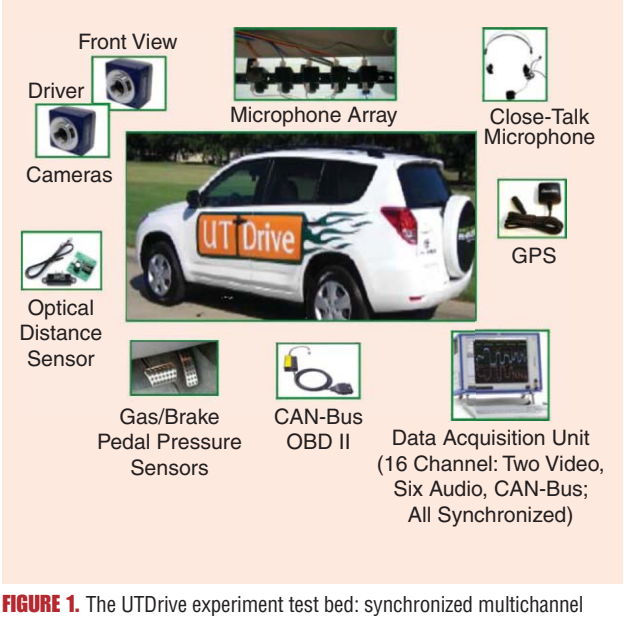


FIGURE 1. The UTDrive experiment test bed: synchronized multichannel measurements. GPS: global positioning system. OBD: on-board diagnostic.

Driver inattention is defined as insufficient or no attention given to activities critical for safe driving. Inattention can either be a voluntary or involuntary diversion of attention by the driver [15]. Driver distraction has been formally defined as "[a]nything that delays the recognition of information necessary to safely maintain the lateral and longitudinal control of the vehicle (primary driving task) due to some event, activity, object or person, within or outside the vehicle (agent) that compels or tends to induce the driver's shifting attention away from the fundamental driving task (mechanism) by compromising the driver's auditory, biomechanical, cognitive or visual faculties or combinations thereof (type)" [16]. Without these formal definitions, cross-study comparisons cannot be made and statistics can vary drastically, leading to incorrect observations [15], [16]. It is important to note that driver distractions are generally caused by a competing trigger activity that may lead to driver inattention, which in turn degrades driving performance. Alternatively, other forms of driver inattention might not necessarily be due to a trigger or competing activity, making inattention difficult to detect and even harder to control. By identifying some of the causes of driver distraction, it is possible to isolate scenarios when the cause of distraction can be controlled.

Most secondary tasks are not distracting and do not require the complete attention of the driver. However, while executing a complex task such as driving, most the driver's attention is toward a safe drive, and performing a secondary task means sharing limited available human cognitive resources. Some important characteristics related to secondary tasks that distract the driver include the duration of the activity, the frequency of the activity, the attention required to execute the activity (attention demand), the ease of returning to the primary task of driving, the location and time at which the activity is executed, and the individual driver's comfort in executing the task and in performing multiple tasks. Since visual modality has been well studied, it has been established that diversion of the driver's visual focus away from the task of driving for more than 1.5 s distracts the driver [17].

The driver follows road rules and maintains his or her lane as well as an acceptable gap between the car and surrounding vehicles, all while achieving good reaction time to changes such as traffic signs and taillights [18]. From the vehicle control side, the driver's primary physical contacts are the steering wheel, the gas and brake pedals, the seat, and the ego vehicle (i.e., the targeted, controlling vehicle itself) speed as reference. Any secondary task that distracts the driver has a direct influence on body movements that manifest in control of the vehicle. Hence, a change in driving performance can be evaluated by analyzing these signals. Each driver has a comfortable way

in which he or she interacts with the vehicle, and analyzing these signals can help build a driver behavior and characteristic model.

Maneuver recognition and distraction analysis

The ability to continuously evaluate driving performance will be necessary in next-generation smart vehicles, to develop advanced driver-specific active/passive safety systems. One typical approach is to identify careless and risky driving events through analyzing abrupt variations in vehicle dynamics information. These variations are best captured when evaluated against similar driving patterns or maneuvers. This has been predominantly adopted in current-day active safety systems [19]–[21]. These event detection systems provide an insight into the current driving conditions of the driver. In addition, every driver has his or her own unique style of driving. Along with weather and traffic, the driver's driving experience, vehicle handling ability, and mental and physical state all influence the way a maneuver is executed. Figure 2 depicts a system in which the driver is identified based on his or her driving characteristics; the driver's maneuvers are recognized, variations in them are identified, and the driving is thus classified. The driver identification subsystem reduces the variability for individual drivers, which can be achieved from face/speech recognition and other inputs. Next, the driving performance is evaluated by identifying maneuvers and detecting their variations against regular (normal execution) patterns. Finally, every driving instance (i.e., in terms of processing frames) is classified into neutral (normal driving) or distracted driving. This section is focused on the maneuver recognition, variation detection, and driving classification subsystems for the distraction analysis.

The ability to continuously evaluate driving performance will be necessary in next-generation smart vehicles, to develop advanced driver-specific active/passive safety systems.

With the pending availability of a massive free-style naturalistic driving data corpus (i.e., Strategic Highway Research Program 2 and New Energy and Industrial Technology Development [22], [23]), the development of automatic tools to organize, prune, and cluster drivercentric-based events for driver modeling is a growing research topic. Rather than using simulated or fixed test track data, it is important to analyze on-road, real-traffic naturalistic driving data for all possible driving variations in different maneuvers.

Human transcription of these massive corpora is not only a tedious task, but also subjective and prone to errors. These human transcription errors can potentially hinder the development of algorithms for advanced safety systems and lead to performance degradations. Therefore, an automatic, effective, and computationally efficient tool is needed to help mitigate human transcription errors and make valuable data from large naturalistic driving corpora more accessible. To prevent these errors from propagating, an automatic maneuver activity detection (MAD) tool (that also detects boundaries) using filter-bank analysis of vehicle dynamic signals is proposed. Using a minimal set of generic vehicle dynamic sensor information, such a MAD tool can match human transcription to an accuracy of up to 99% [24], [25]. Making this tool freely available will offer researchers opportunities to better explore naturalistic driving data.

Maneuver recognition

Driving maneuvers, influenced by the driver's choice and traffic/road conditions, are important in understanding variations in driving performance and to help rebuild the intended route. Maneuvers are the basic units in building up a driving session. While processing massive quantities of naturalistic driving data, it is critical to analyze at a micro level. Understanding how these maneuvers are performed can provide information on how the driver controls the vehicle and how driving performance varies over time, which is essential in driver assistance and safety systems.

Similar to speech, where phonemes form words, it has been established [26], [27] that the smallest meaningful units of a driving pattern are termed *drivemes*. Drivemes form maneuvers, and maneuver sequences form a navigation route. This flow is depicted in Figure 3. Therefore, tracking the variation of these drivemes can improve the efficiency of active safety systems not only in providing safety to the driver, but also in predicting drivers' actions.

The definition of driving maneuvers may be considerably wide, depending on the underlying application [28]. Several existing studies have employed maneuver recognition for vehicle trajectory prediction [29], intersection assistance [30], and lane-change intent recognition on the highway [31]. Based on recent advancements, a study [32] that considered driving maneuvers primarily classified into eight categories—straight, stop, left turn, right turn, left-lane change,

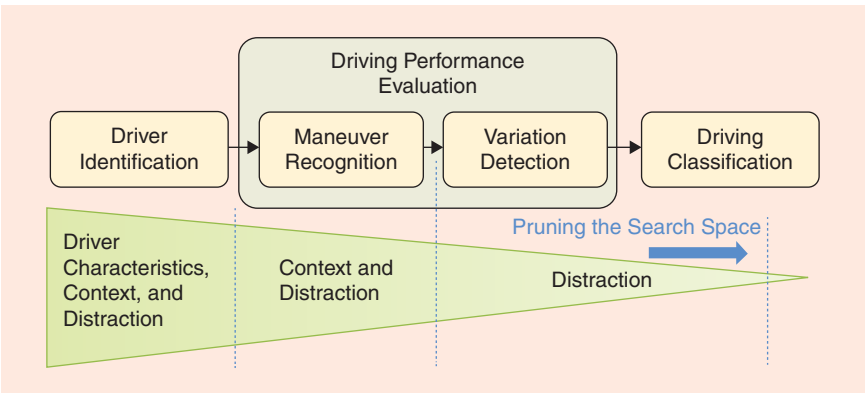


FIGURE 2. How the driver-dependent, maneuver-based distraction detection system identifies and evaluates each driving session.

right-lane change, left road curve, and right road curve—showed promise.

The method of recognition in the literature employed various statistical modeling and machine-learning classification algorithms, such as Bayesian models [33], finite-state machines and fuzzy logic [34], hidden Markov models (HMMs) [35], and decision trees [36]. HMMs have proven to be beneficial in predicting driver actions within the first 2 s of an action sequence [37]. In our previous study, a similar HMM framework was employed in both a top-down as well as bottom-up approach to find the best integrated architecture for modeling driving behavior and recognizing maneuvers and routes [38]. Important features include steering wheel angle, speed, and brake signals from vehicle CAN bus data, or acceleration and gyroscope readings from a smart portable device [25], [39]. Recognition and prediction of lane-change maneuvers have been proposed together, suggesting a double-layered HMM framework in the consideration of both maneuver execution and route information [40]. Thus far, the accuracy of obtained maneuver recognition ranges between 70–90% and offers opportunities for low-cost, low-level maneuver recognition for long-term modeling of driver behavior.

Rather than using simulated or fixed test track data, it is important to analyze on-road, real-traffic naturalistic driving data for all possible driving variations in different maneuvers.

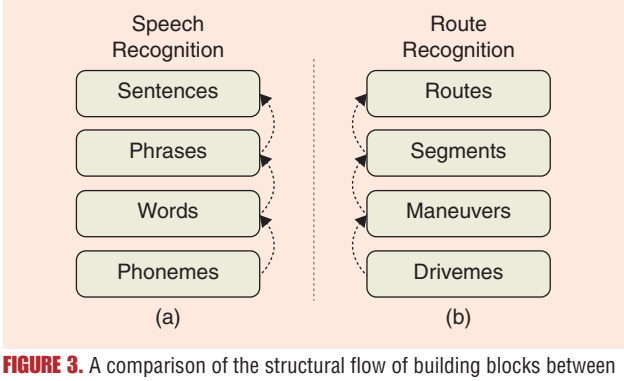


FIGURE 3. A comparison of the structural flow of building blocks between (a) speech recognition and (b) route recognition.

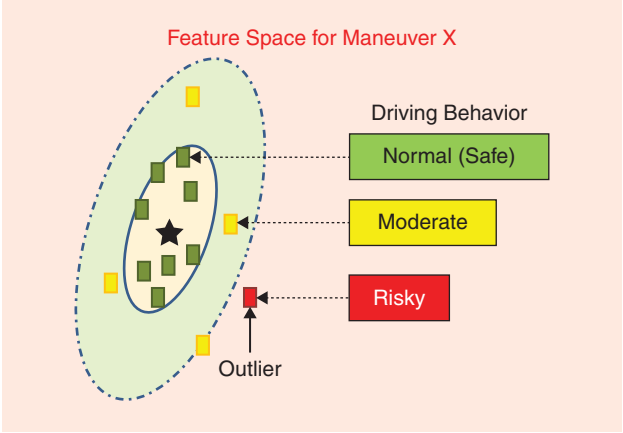


FIGURE 4. An example of the feature space for maneuver X showing variations in driving performance quantified as normal, moderate, and risky maneuver actions.

Distraction analysis

Distraction, in general, affects the attention span of a person; within the vehicular space, it manifests in the driver's vehicular controls. Traditionally, distraction has been assessed from the driver's perspective in terms of either stress, eye movements, or cognitive workload [41]. Physiological measurements such as heart rate variability and skin conductance (e.g., electroencephalogram, electromyogram) have proven to be useful in detecting the stress levels in drivers [42]. Studies have also considered body movement sensors to detect drivers' patterns for assessment of driver distractions [43]. Though high accuracy has been achieved from a research perspective, these vision and body sensors are intrusive and unsuitable for naturalistic driving scenarios. Using such sensors can potentially serve as a baseline when compared with nonintrusive sensors for performance.

Since driver actions and intentions manifest into vehicle movement, vehicle dynamic signals such as steering wheel movements, gas and brake pedal pressure, and vehicle speed could potentially contain hidden or embedded information on the current status of the driver. Using vehicle dynamic signals, driving is classified based on the maneuver execution characteristics of a particular driver. The classification could be a binary classification (neutral versus distracted) [46] or a trend in the variations (safe, moderate, or risky) [32].

The assessment of driving distraction underlies two hypotheses. First, good, safe, or convenient driving behavior should be reflected by stable, steady vehicular dynamic performance. Second, the actions of an experienced driver should meet the characteristics of good driving behavior most of the time, with bad driving occurring as a limited number of events [44]. Based on these hypotheses, the good driving events should be clustered in the vehicle dynamical feature space, whereas bad driving events will become more random anomalies or outliers.

Figure 4 depicts a typical feature space for an imaginary maneuver type, X. The green squares, which are clustered together around the centroid of class X, represent the normal execution trend for this maneuver. The deviations from the normal execution pattern are reflected in the feature space of this maneuver as yellow or red squares. These abnormal instances of the maneuver are still recognized as type X by the classifier, but the intraclass separation suggests that they can be marked as outliers. Euclidean distance, cosine distance, and Mahalanobis distance have been used to detect outliers. Identifying such outliers helps in the evaluation of driving pattern variations and driving performance [45]. Figure 5 illustrates the gradient of event variations (classified as safe, moderate, and risky) along the driving route.

Due to the highly dynamic nature of driving and the surrounding environment, drivers generally do not stay in one state for long and often toggle between models/states. A microanalysis of individual driving patterns is performed by segmenting the drive into small frames (a few seconds or a few meters traveled), which can be scaled to a macro level for preventing or correcting any unsafe activities. Such a microanalysis has provided an insight into how secondary tasks are executed and potentially influence drivers. Most secondary tasks can be grouped into three sequential events [46]. In the anticipation/preparatory phase, during the start of a task, most drivers are distracted. This is justified as they divert more attention toward the task, assess the surroundings, and get ready to perform the task. The second event is the task execution phase, during which the drivers fall into a comfort zone of multitasking. Finally, in the third task, the recovery or postcompletion phase, drivers generally reassess their surroundings after secondary task completion. The duration of each of these phases is based on the individual driver's comfort and confidence level [47]; the effect of multitasking is variable on different drivers. As the automotive industry further advances in developing advanced driver-assistance systems (ADASs), such drivercentric adaptive systems will help in personalizing the vehicle by triggering the ADAS only when drivers are impacted or when they show tendencies of such impact.

The National Highway Traffic Safety Administration (NHTSA) released visual driver distraction guidelines [17] for in-vehicle electronic devices, categorizing the main sources of distraction into three categories: visual, cognitive, and manual. It will not be long before the automotive industry and infotainment systems shift away from visual interaction with the driver and move toward audio/speech-based interactions with the driver. Therefore, it is of great interest to understand the actual influence of in-vehicle speech on the driver. There has been some preliminary work done in this area to understand the influence of in-vehicular speech and audio on driving [48]. While some in-vehicle conversations might aid driving, categories such as involved, competitive, and argumentative speech can adversely influence the driver and cause driver distraction.

Due to the highly dynamic nature of driving and the surrounding environment, drivers generally do not stay in one state for long and often toggle between models/states.

Tracking glance behavior and visual attention

An important aspect in monitoring driver distraction is to evaluate the visual attention of the driver. There are three main areas that can benefit from tracking the drivers' visual attention: assessing the primary driving task, detecting secondary tasks, and supporting advanced user–computer interfaces.

The role of visual attention

Understanding where the visual focus lies is a key step to determine driver performance during the primary driving task [49]–[51]. A driver should scan the route environment before conducting a driving maneuver. This action includes checking the mirrors, looking at the vehicles in front of the driver, and

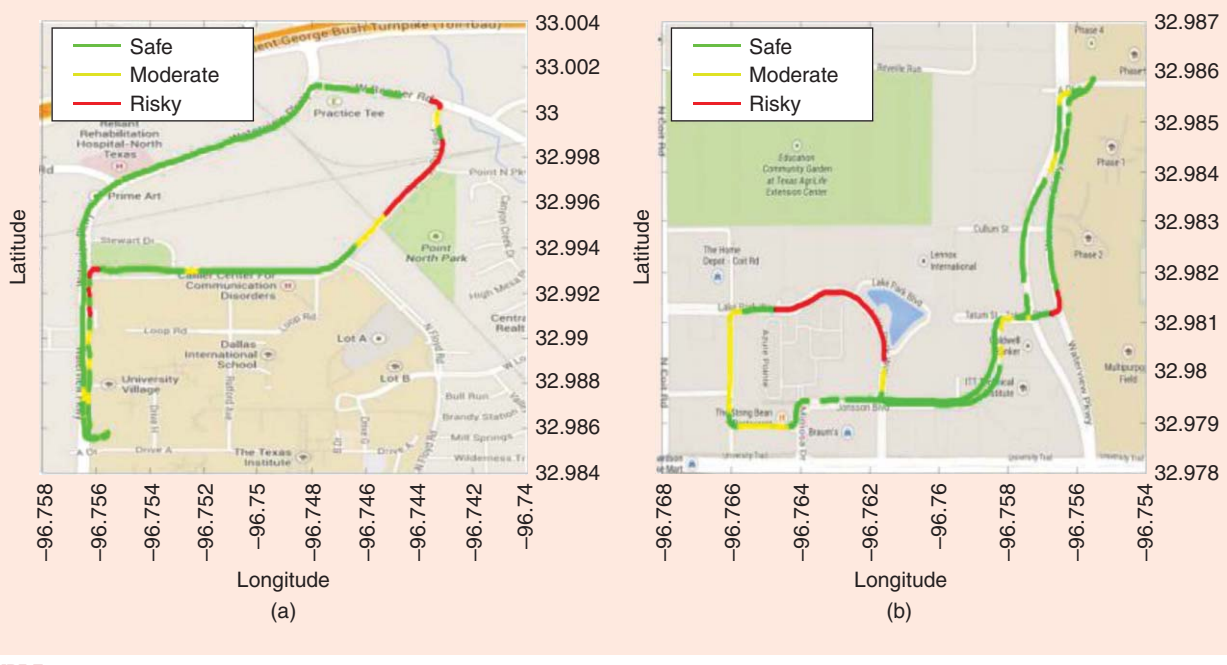


FIGURE 5. (a) and (b) This chart shows maneuver variations along two different driving routes. The routes were selected around the University of Texas at Dallas campus, with a mixture of residential and business areas. (a) and (b) The green, yellow, and red colors indicate driving safety levels along the two routes.

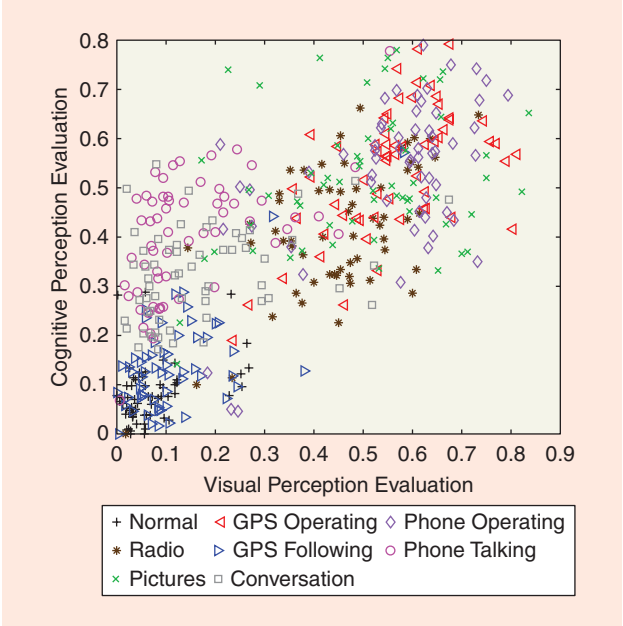


FIGURE 6. An overview of the visual–cognitive space proposed to study driver distraction. Perceptual evaluations are used to determine the perceived level of distractions. Each secondary task affects the driver's concentration differently.

identifying pedestrian actions. Primary driving tasks such as visual scanning, turning, and switching lanes all require mirror-checking actions [52]–[54]. Failing to accomplish these tasks decreases the drivers' situational awareness, increasing the chances of accidents [55], [56]. An increase in visual demand due to secondary tasks affects the control of the vehicle, the detection of critical traffic events, and the detection of hazard events [57]. As a result, studies have used features describing eye-off-the-road, head pose, gaze range, and eyelid movements to detect distractions [58]–[65]. Objective measures capturing the duration and frequency of glance behaviors can provide important information to provide warnings to distracted drivers.

Visual attention signals temporal deviations from the primary driving task to complete secondary tasks such as adjusting the radio, operating a cell phone, or looking at other passengers. All of these secondary tasks induce visual, cognitive, auditory, and manual distractions. A perceptual evaluation has been conducted to assess the perceived level of cognitive and visual distractions in 10-s videos of drivers who are engaged in different secondary tasks [65], [66], in which the advantages and limitation of using perceptual evaluations to assess driver distractions is discussed. Figure 6 shows that many common secondary tasks induce a high level of visual distractions. For example, operating a cell phone, the radio, or a navigation system increases the perceived level of visual distractions [67]. When a driver fails to glance at traffic, it can also signal cognitive distractions; the driver is looking but not seeing because he or she is daydreaming or thinking about something else [68], [69]. These types of distractions are very difficult to detect with noninvasive sensors [70]. Tracking glance behaviors provides an important tool to address this problem. For all of these reasons, a robust ADAS should be able to detect mirror-checking actions and glance behaviors to prevent hazard situations [71].

The automobile industry is developing new advanced interfaces that do not induce manual or visual distractions. These interfaces are generally implemented using automatic speech-recognition systems. ADASs need to provide essential information to the driver in an effective manner. With more information available to the driver, it is also important that the information is presented without causing significant distractions. By tracking the visual attention of the driver and environment, the ADAS can clarify ambiguities by providing a situated dialog system (e.g., commands such as "What is the address of this building?" while glancing toward a specific building). In an example of such a system [72], the visual saliency of the scene and crowdsourced statistics on how people describe objects were used as prior information to improve the identification of points of interest (POIs). While the visual saliency of the scene did not depend on driver glance behaviors, we expect improved performance by modeling the visual attention of the drivers [73], [78].

Tracking visual attention

Tracking eye movement can be an accurate measurement to identify the exact location of the gaze of the driver. However, robustly measuring gaze in a driving environment is challenging due to changes in illuminations in the vehicle and changes in the head poses of the drivers. As a result, most of the studies have approximated gaze with head poses. Zhang et al. [74] argued that even though eye gaze is a better indicator, head pose alone can provide good cues about driver intentions. However, there are differences between head pose and the driver's gaze that need to be considered [75]–[78]. The driver moves his or her head and eyes to glance at a target object, where the eye–head relationship depends on factors such as the underlying driving task, the type of road, and the driver.

Studies have investigated the relation between head motion and gaze on naturalistic recordings [78]. We placed multiple markers on the windshield, side windows, speedometer panel, radio, and gear. The recordings protocol is repeated while driving and when the car was parked. We proposed regression models where the dependent variables were the position and rotation of the head, and the independent variables were the three-dimensional positions of the POIs. While driving, the R^2 of the model was about 0.73 for the horizontal direction, but lower than 0.20 for other directions. Motivated by these results, the analysis is extended to incorporate a probabilistic model relying on Gaussian process regression [79]. Instead of deriving the exact location of the POI, the framework creates a salient visual map describing the driver's visual attention, which is mapped into the route scene (see Figure 7). The 95% confidence region of the models included about 89% of the POIs. This approach provides a suitable tool for situated dialog systems and safety systems that are aware of the driver glance behavior.

An alternative approach to monitoring visual attention is to directly recognize primary or secondary driving tasks that require visual demand. An example of a primary driving task is the detection of mirror-checking actions. We presented an accurate random undersampling boost classifier to recognize mirror-checking actions [71]. The classifier was trained with multimodal features automatically extracted from the driver and road cameras and from the CAN bus signal using naturalistic recording on the UTDrive platform. The task was to recognize each time the driver looked at a given mirror. Figure 8(a) shows an example of a participant looking at the rear mirror. The F-score of the classifier was 91.4%, which is very high given that mirror-checking actions are infrequent events, making this classification problem highly unbalanced. Figure 8(b) shows the performance for different routes, under both normal conditions (during which the driver is not engaged in secondary tasks) and task conditions (during which the driver is engaged in secondary tasks). The classifier showed consistent performance across both normal and task conditions. An example of a secondary task is the detection of activities not related to the driving task requiring visual activities. We trained binary classifiers using a support vector machine, which detect particular secondary activities [60]. For tasks such as looking at pictures (which simulates the task of looking at billboards, sign boards, and shops) and operating a radio and a GPS, the accuracy was about 80%. The perceptual evaluation showed that these tasks induce high visual demand.

Portable platform advancements

With the rapid growth of smartphone capabilities, including entertainment and management of daily activities, individuals are increasingly using smartphones while driving. However, operating a vehicle is a complicated and skilled task requiring multimodal (especially visual) attention and focus. While drivers can multitask comfortably, using a smartphone may become a distraction and contribute to increased risk. Alternatively, the proper use of smart devices could be a source of reduced driving distraction while executing secondary tasks. One feasible approach to achieve this is to interact with these platforms using speech-based interfaces, reducing the visual and cognitive load [6]. Studies have also shown that drivers can achieve better and safer driving performance while using speech interactive systems to operate in-vehicle systems compared to hand-operated interfaces [80].

A more advanced reason for introducing the smartphone is its potential ability to be integrated with intelligent telematics services. Smartphone-based on-board sensing in the vehicle can capture various sources of information, including traffic (other vehicle and pedestrian movements), vehicle (diagnostics), environment (road and weather), and driver behavior information [81]. It would be beneficial to connect this platform

FIGURE 7. This visual saliency map was created with the probabilistic model and shows (a) the estimation of confidence regions for different distances from the car and (b) an aggregation of the results projected on the road camera. The saliency map characterizes driver visual attention.

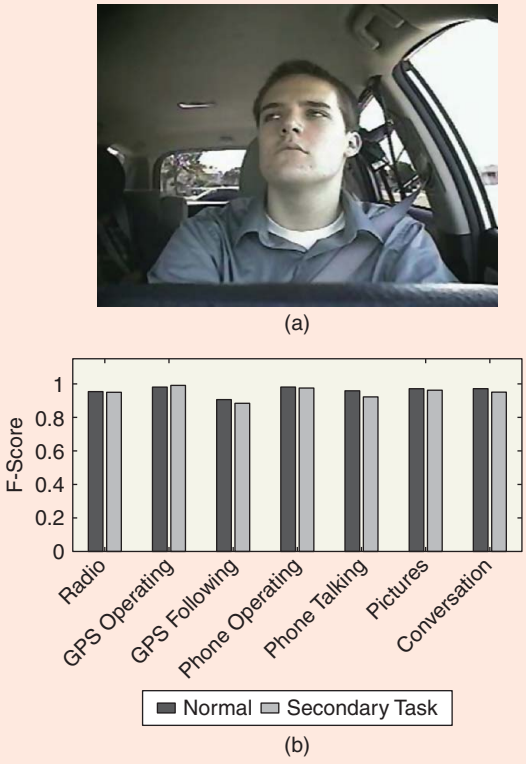


FIGURE 8. The detection of mirror checking using multimodal features: (a) a participant looking at the rear mirror and (b) the F-score of the classifier for both normal and task conditions [71].

with intelligent transportation systems or V2V or V2I communication, share the information, and realize a wider Internet of Vehicles. However, the challenges of smartphone platform use in the vehicle come from the deployment difficulty, measurement accuracy, and system reliability.

In this section, we first discuss the deployment of smartphones as an in-vehicle data collection platform, utilizing its hardware resources. Additionally, we explore the implementation of the voice-based human–machine interface, and its capabilities in applications for vehicle/driver telematics.

In-vehicle data collection platform: Mobile-UTDrive app

Smartphones contain a variety of useful sensors, including cameras, microphones, inertial measurement units (IMU), and GPS. These multichannel signals make the smartphone a potentially leveraged platform for in-vehicle data sensing and monitoring, and can be employed for driving distraction analysis. The use of smart portable devices in vehicles creates the possibility to record useful data and helps develop a better understanding of driving behavior. This option allows a wider range of naturalistic driving study opportunities for drivers operating their own vehicles [82]–[84].

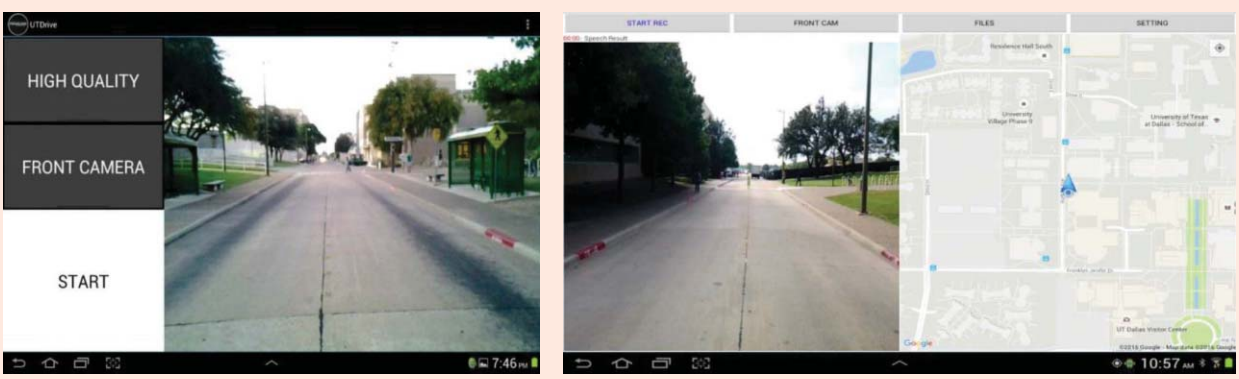
Freely distributing this platform will offer researchers the opportunity to customize their data collection scenarios while maintaining current goals for naturalistic driving data advancements.

The UTDrive mobile app (Mobile-UTDrive) has been developed with the goal of improving driver/passenger safety while simultaneously maintaining the ability to establish monitoring techniques that can be used on mobile devices in various vehicles [85]. Mobile-UTDrive has been primarily used and developed as a multimodal data acquisition platform that collects driver, vehicle, and environmental information describing the comprehensive driving scenario. The modalities captured by Mobile-UTDrive are audio, video, GPS, and IMU sensor signals. The app runs on any Android-based smart portable device and uses the front and rear cameras to record naturalistic driving video as well as in-vehicle audio. The IMU and GPS within the device provide accurate estimates of vehicle dynamics. Mobile-UTDrive has been further developed to take advantage of capabilities such as speech recognition and on-screen map navigation. Figure 9 displays a screenshot of the Mobile-UTDrive app running on a tablet. Using this approach has resulted in studies to detect maneuvers and design driving safety systems that combine in-vehicle speech and video analysis with driving performance evaluation [86], [87]. Freely distributing this platform will offer researchers the opportunity to customize their data collection scenarios while maintaining current goals for naturalistic driving data advancements.

In previous studies, it has been shown how vehicle dynamic signals can potentially replace the information extracted from a CAN bus and thereby extend the use of maneuver recognition and monitoring algorithms to any vehicle that uses the app [32], [45]. However, the orientation and relative movement of the smartphone inside the vehicle yields the main challenge for platform deployment. A recent study [88] proposed a solution of converting the smartphone-referenced IMU readings into vehicle-referenced accelerations, which allows free positioning of smartphones for the in-vehicle dynamics sensing. In this proposed framework, the raw smartphone IMU readings are first processed through a geometry coordinate transformation to rotate/reorient the smartphone-referenced accelerations into a vehicle-referenced coordinate system. Next, a regression model is established to map the relationship between IMU and GPS data, and therefore provide an adaptive filtering process to decouple the smartphone's relative movement in the vehicle. This serves as a preprocess module and therefore provides the basis for further applications using the smartphone data (see Figure 10).

Voice-based human–machine interaction

The smartphone in the vehicle provides easy access to speech recording and processing, and offers potential integration with



(a) (b)

FIGURE 9. The mobile-UTDrive app display showing (a) the original and (b) the updated version [85].

the infotainment system, which becomes a good platform for the development of a voice-based human–machine interface. Drivers would not necessarily be required to perform tasks such as setting map navigation, changing the radio station, adjusting volume, adjusting the air-conditioning, and controlling the windows via hands-on operation, but could employ voice commands instead. The typical manual-entry or tactile-based engagement primarily uses various combinations of keypads, keyboards, point-and-click techniques, touch-screen displays, or other interface mechanisms. These traditional interfaces, which often require the driver to take his or her eyes off the road, tend to be cumbersome in environments where the speed of interactions and dangers of distraction pose significant issues, and therefore fall short in providing simple and intuitive operation. In contrast, voice-based interaction would keep the driver's eyes focused on the road and his or her hands on the wheel. Results have shown that hands-on operation could potentially be a greater cause of major irregularities in driving performance, despite the latency and recognition error imposed by the speech recognition system [45], [87]. The development of speech recognition compatibility and natural language understanding for dialog interaction in the car offers avenues for lowering driver distraction. Therefore, natural voice-based engagements between driver and vehicle offer the potential to meet an ever-growing demand for creating a comfortable, safe, and convenient driving experience.

Among the voice-based human–vehicle interfaces, the navigation dialogue system is the one with the highest demand in recent years. Navigation dialogs may happen in some situations while a user may be driving, on-the-go, or in other environments where having a hands-free interface provides critical advantages. The desired intelligent navigation system is about more than searching locations on the map; it would have the capability of acting as an assistant, talk with humans in a natural manner, and guide and drive for the human when needed. Therefore, it should have the ability to speak naturally and understand natural spoken language. For example, when a driver is trying to find a destination, he or she may either speak out a POI, specify the exact address, or spell the name and number of a street. The navigation system should automatically understand what was said without having to ask the driver to choose the style of his or her spoken language. Furthermore, in the next generation of autonomous driving vehicles, it is expected that the vehicle will automatically drive for the human. Passengers may inquire about the trip or change the previously selected route through the dialog system, and the vehicle should be able to understand how it is associated with navigation tasks and provide the necessary responses.

Recent studies [89], [90] consider natural language processing (NLP) for the navigation-oriented human–vehicle dialog. The NLP framework is based on a recurrent neural network and long short-term memory architecture, and contains sentence-level sentiment analysis and word/phrase-level context extraction. As shown in Figure 11, the sentiment analysis identifies whether a sentence is navigation related. If the sentence is navigation related, the next stage is to extract useful context by recognizing the word/phrase labels. The extracted information will be ready to submit for response or path planning. The accuracy of NLP was 70–98%, depending on the accuracy of the speech recognition results.

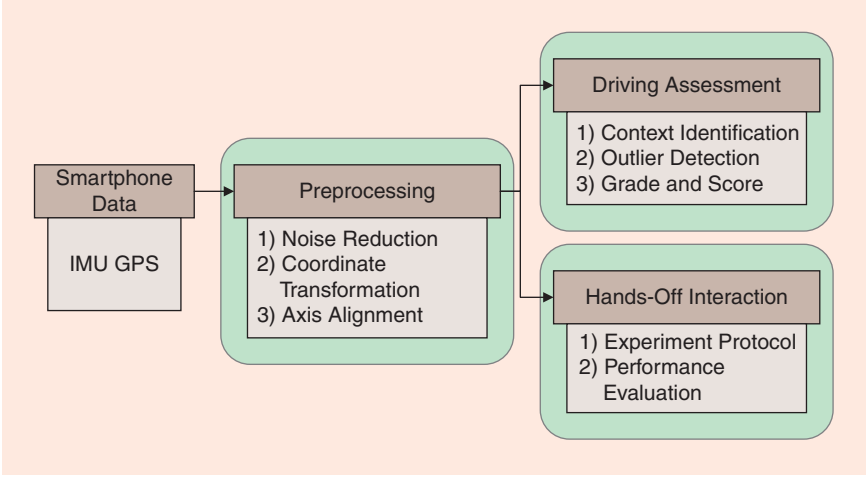


FIGURE 10. The smartphone data processing modules.

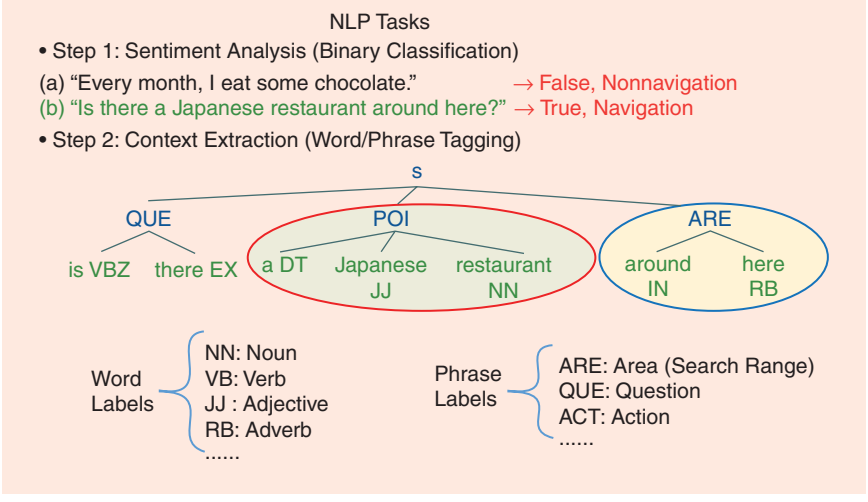


FIGURE 11. An example of the processing for NLP tasks. For the sentiment analysis of two sentences, (a) is not navigation related, while (b) is navigation related. Therefore, (b) is further processed with context extraction, and useful information (such as POI and search area) is labeled.

Discussion and conclusions

Vehicle technologies have advanced significantly in terms of improved transportation, comfort, and safety, and will continue to evolve as we move forward into the next generation of transportation systems and infrastructure. From this article, as well as the broad coverage from recent articles in the November 2016 issue of *IEEE Signal Processing Magazine,* it is clear that new technologies are migrating into novel in-vehicle systems for route navigation, information access, infotainment, and connected vehicle advancements for V2V and V2I connectivity and communications. Vehicle driving autonomy is also evolving, and further research advancements are needed to better understand the interplay between the driver, the vehicle, and the route/environment. With the motivation of contributing to improved intelligent driver–vehicle systems that incorporate human-specific characteristics, the CRSS-UTDrive Lab has focused their research on naturalistic driving studies, with the interest of understanding driver behavior and distraction from multichannel sensor data. Any secondary driver task activity in the vehicle can be a source of driving distraction and therefore impact driving performance. Regarding this, one typical approach is to first extract the driving context in terms of microlevel components (e.g., maneuvers), and then evaluate risky events and variations against similar driving patterns in the vehicle dynamics domain. An alternative approach is to directly monitor drivers' physical or glance behavior and assess their cognitive and visual attention. Previous studies have shown precise results in the detection of driving distraction, driving performance analysis, and visual attention tracking. To take advantage of the fast-growing smartphone applications market and integrate telematics services, recent activities have resulted in a mobile platform that contributes to in-vehicle naturalistic driving studies and voice-based human–machine interfaces. These studies, if combined, would be able to provide a comprehensive understanding of the driver's state and driving performance, establish a comfortable driving experience with a humancentric assistant in the vehicle, and contribute intelligent transportation information sharing via V2V/V2I connectivity.

While there is great interest in migrating to fully automated, self-driving vehicles, next-generation vehicles will need to be more active in assessing driver awareness, vehicle capabilities, traffic/environmental settings, and how these factors come together to determine a collaborative, safe, and effective driver–vehicle engagement for vehicle operation. Greater interdisciplinary research that addresses multifunctional vehicles to support smooth transitions from complete human control toward semisupervised/assisted control and even fully automated scenarios are needed. In the end, while signal processing technical advancements can enhance vehicles and the comfort, enjoyment, and capabilities of drivers, to be successful these efforts must first do no harm and ensure improved safety.

Authors

John H.L. Hansen (john.hansen@utdallas.edu) received his B.S.E.E. degree from Rutgers University, New Jersey, and his M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, respectively. He serves as associate dean for research at the University of Texas at Dallas, where he founded the Center for Robust Speech Systems–UTDrive Lab and the Robust Audio and Speech Processing Lab. He has served as a technical program chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing 2010, the general chair for the International Speech Communication Association (ISCA) INTERSPEECH 2002, and as an organizer/contributor for the Biennial Workshop on Digital Signal Processing for In-Vehicle Systems and Safety (2003–2015), as well as the corresponding edited textbook series on driver modeling and safety. He is an ISCA fellow, IEEE Fellow, and an Acoustical Society of America's 25-Year Award recipient. He has coauthored more than 650 papers in the fields of digital signal processing, speech processing, machine learning, and driver distraction modeling/detection.

Carlos Busso (busso@utdallas.edu) received his B.S. in electrical engineering from the University of Chile in 2000, his engineering degree in electrical engineering from the University of Chile in 2003, his M.S. degree in electrical engineering from the University of Chile in 2003, and his Ph.D. degree in electrical engineering from the University of Southern California in 2008. Currently, he is an associate professor at the University of Texas at Dallas, where he oversees the Multimodal Signal Processing Lab. His current research includes the broad areas of in-vehicle modeling of driver behavior, affective computing, and multimodal human–machine interfaces. He is a recipient of the National Science Foundation CAREER Award, the International Conference on Multimodal Interaction Ten-Year Technical Impact Award, and the Hewlett Packard Best Paper Award at the IEEE International Conference on Multimedia and Expo 2011 (with J. Jain). He is an associate editor of *IEEE/ACM Transactions on Audio, Speech, and Language*. He is a Senior Member of the IEEE.

Yang Zheng (yxz131331@utdallas.edu) received his B.S. and M.S. degrees in automotive engineering from Nanjing University of Aeronautics and Astronautics, China, and Tongji University, Shanghai, China, in 2010 and 2013, respectively. He is currently a Ph.D. student in electrical engineering at the University of Texas at Dallas, where he works in the Center for Robust Speech Systems–UTDrive Lab. His research interests include driver behavior analysis, human–machine interface, and intelligent vehicle advancements. He has authored/coauthored three journal articles and 15 conference papers; he also holds two patents. He has served as a reviewer for the IEEE Intelligent Transportation Systems Society and the IEEE Vehicular Technology Society. He is a Student Member of the IEEE.

Amardeep Sathyanarayana (amardeep@utdallas.edu) received his B.E. degree from Visvesvaraya Technological University, India, in 2004 and his M.S. degree in electrical engineering and his Ph.D. degree in signal processing from the University of Texas at Dallas in 2008 and 2013, respectively. He is currently with Uhnder Inc., furthering next-generation

active safety systems. Previously, he worked for Texas Instruments' Kilby Labs, AT&T Shannon Research Labs, Ford, and Bosch. His research interests include active safety systems and statistical signal processing. He has authored more than 30 book chapters, journal papers, and conference papers, and he has filed ten patents. He has served as a reviewer for various journals and conferences, including *Journal of Selected Topics in Signal Processing* and the IEEE International Conference on Intelligent Transportation Systems.

References

[1] J. F. Antin (2011). Design of the in-vehicle driving behavior and crash risk study: In support of the SHRP 2 naturalistic driving study. U.S. Transportation Res. Board. Washington, D.C. Rep. S2-S05-RR-1. [Online]. Available: http://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=2127

[2] J. H. L. Hansen, K. Takeda, S. M. Naik, M. M. Trivedi, G. U. Schmidt, and Y. J. Chen, "Signal processing for smart vehicle technologies," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 12–13, Nov. 2016.

[3] C. Miyajima and K. Takeda, "Driver-behavior modeling using on-road driving data: a new application for behavior signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 14–21, Nov. 2016.

[4] Y. Choi, S. I. Han, S.-H. Kong, and H. Ko, "Driver status monitoring systems for smart vehicles using physiological sensors," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 22–34, Nov. 2016.

[5] A. S. Aghaei, B. Donmez, C. C. Liu, D. He, G. Liu, K. N. Plataniotis, H.-Y. W. Chen, and Z. Sojoudi, "Smart driver monitoring: When signal processing meets human factors," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 35–48, Nov. 2016.

[6] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, and J. H. L. Hansen, "Conversational in-vehicle dialog systems," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 49–60, Nov. 2016.

[7] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, "Recent advances in active noise control inside automobile cabins," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 61–73, Nov. 2016.

[8] R. Hult, G. R. Campos, E. Steinmetz, L. Hammarstrand, P. Falcone, and H. Wymeersch, "Coordination of cooperative autonomous vehicles," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 74–84, Nov. 2016.

[9] N. AbuAli and H. Abou-zeid (2016, Nov.). Driver behavior modeling: Developments and future directions. *Int. J. Veh. Technol.* [Online]. Available: http://dx.doi.org/10.1155/2016/6952791

[10] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich, and A. Schmidt, "A data set of real world driving to assess driver workload," in *Proc. 5th Int. Conf. Automotive User Interfaces and Interactive Vehicular Applications,* 2013, pp. 150–157.

[11] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data," Nat. Highway Traffic Safety Admin., Washington, DC, Rep. DOT HS 810 594, 2006.

[12] UTDrive Corpus Classical. [Online]. Available: http://www.utdallas.edu/research/utdrive/corpus.html

[13] J. H. L. Hansen, P. Boyraz, K. Takeda, and H. Abut, *Digital Signal Processing for In-Vehicle Systems and Safety.* New York: Springer, 2012.

[14] G. Schmidt, H. Abut, K. Takeda, and J. H. L. Hansen, *Smart Mobile In-Vehicle Systems—Next Generation Advancements.* New York: Springer Publishing, 2014.

[15] M. A. Regan, C. Hallett, and C. P. Gordon, "Driver distraction and driver inattention: definition, relationship and taxonomy," *Accident Anal. Prev.*, vol. 43, no. 5, pp. 1771–1781, Sept. 2011.

[16] M. Pettitt, G. Burnett, and A. Stevens (2005), "Defining driver distraction," in *Proc. 12th World Congr. Intelligent Transport Systems*, San Francisco, CA, Nov. 2005, pp. 1–12.

[17] Department of Transportation, "Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices," Nat. Highway Traffic Safety Admin., Washington, DC, Tech. Rep. Docket No. NHTSA-2010-0053, 2012.

[18] K. Young, M. Regan, and M. Hammer, "Driver distraction: a review of the literature," Accident Res. Centre, Monash Univ., Melbourne, Victoria, Tech. Rep. No. 206, 2003.

[19] S. Boonmee and P. Tangamchit, "Portable reckless driving detection system," in *Proc. 6th Int. Conf. Electrical Engineering/Electronics Computer Telecommunications and Information Technology*, Bangkok, Thailand, 2009, pp. 412–415.

[20] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *Proc. 4th Int. Conf. Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, Munich, Germany, 2010, pp. 1–8.

[21] D. A. Johnson and M.M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. 14th Int. IEEE Conf. Intelligent Transportation Systems*, Washington, D.C., 2011, pp. 1609–1615.

[22] K. Takeda, J. H. L. Hansen, H. Erdogan, and H. Abut, *In-Vehicle Corpus and Signal Processing for Driver Behavior.* New York: Springer, 2009.

[23] K. Takeda, J. H. L. Hansen, P. Boyraz, L. Malta, C. Miyajima, and H. Abut, "An international large-scale vehicle corpora for research on driver behavior on the road," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1609–1623, Dec. 2011.

[24] A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, "Automatic maneuver boundary detection system for naturalistic driving massive corpora," *SAE Int. J. Passeng. Cars—Electron. Electr. Syst.*, vol. 7, no. 1, pp. 149–156, Apr. 2014.

[25] A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, "Leveraging sensor information from portable devices towards automatic driving maneuver recognition," in *Proc. 15th Int. IEEE Conf. Intelligent Transportation Systems*, Anchorage, AK, 2012, pp. 660–665.

[26] K. Torkkola, S. Venkatesan, and H. Liu, "Sensor sequence modeling for driving," in *Proc. 18th Int. Florida Artificial Intelligence Research Society Conf.*, Clearwater Beach, FL, 2005, pp. 721–727.

[27] A. Sathyanarayana, P. Boyraz, and J. H. L. Hansen, "Driver behavior analysis and route recognition by hidden Markov models," in *Proc. IEEE Int. Conf. Vehicular Electronics Safety*, Columbus, OH, 2008, pp. 276–281.

[28] A. D'Agostino, A. Saidi, G. Scouarnec, and L. Chen, "Learning-based driving events classification," in *Proc. 16th Int. IEEE Annu. Conf. Intelligent Transportation Systems*, 2013, pp. 1778–1783.

[29] A. Houenou, P. Bonnifait, V. Cherfaoui, and W. Yao, "Vehicle trajectory prediction based on motion model and maneuver recognition," in *Proc. 2013 IEEE/RSJ Int. Conf. Intelligent Robots Systems*, Tokyo, Japan, 2013, pp. 4363–4369.

[30] Q. Tran and J. Firl, "Online maneuver recognition and multimodal trajectory prediction for intersection assistance using non-parametric regression," in *Proc. 2014 IEEE Intelligent Vehicles Symp.*, Dearborn, MI, 2014, pp. 918–923.

[31] J. Nilsson, J. Fredriksson, and E. Coelingh, "Rule-based highway maneuver intention recognition," in *Proc. 2015 IEEE 18th Int. Conf. Intelligent Transportation Systems*, Las Palmas, Spain, 2015, pp. 950–955.

[32] A. Sathyanarayana, O. Sadjadi, and J. H. L. Hansen, "Automatic driving maneuver recognition and analysis using cost effective portable devices," *SAE Int. J. Passeng. Cars—Electron. Electr. Syst.*, vol. 6, no. 2, 2013, pp. 467–477.

[33] A. Gerdes, "Automatic maneuver recognition in the automobile: The fusion of uncertain sensor values using Bayesian models," in *Proc. 3rd Int. Workshop Intelligent Transportation (WIT 2006)*, Hamburg, DE, 2006, pp. 129–133.

[34] T. Hülnhagen, I. Dengler, A. Tamke, T. Dang, and G. Breuel, "Maneuver recognition using probabilistic finite-state machines and fuzzy logic," in *Proc. IEEE Intelligent Vehicles Symp. (IV),* San Diego, CA, 2010, pp. 65–70.

[35] D. Mitrovic, "Reliable method for driving events recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 198–205, June 2005.

[36] Y. Zheng, A. Sathyanarayana, and J. H. L. Hansen, "Threshold based decision-tree for automatic driving maneuver recognition using CAN-Bus signal," in *Proc. IEEE Intelligent Transportation Systems Society Conf.*, Qingdao, China, 2014, pp. 2834–2839.

[37] A. Pentland and A. Liu, "Modeling and prediction of human behavior," *Neural Comput.*, vol. 11, no. 1, pp. 229–242, 1999.

[38] A. Sathyanarayana, "Multi-modal signal processing in-vehicular systems for driver distraction identification and driver behavior modeling," Master's thesis, Dept. Electrical Engineering, Univ. Texas at Dallas, 2008.

[39] Y. Zheng, A. Sathyanarayana, and J. H. L. Hansen, "Non-uniform time window processing of in-vehicle signals for maneuver recognition and route recovery," presented at the SAE World Congr., Detroit, MI, 2015.

[40] Y. Zheng and J. H. L. Hansen, "Lane-change detection from steering signal using spectral segmentation and learning-based classification," *IEEE Trans. Intell. Transp. Syst.,* to be published.

[41] T. B. Hughes, H.-S. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of an HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 346–349, May 1999.

[42] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, June 2005.

[43] A. Sathyanarayana, S. Nageswaren, H. Ghasemzadeh, R. Jafari, and J. H. L. Hansen, "Body sensor networks for driver distraction identification," in Proc. *IEEE Int. Conf. Vehicular Electronics Safety*, Columbus, OH, pp. 120–125, Sept. 2008.

[44] Y. Zheng and J. H. L. Hansen, "Unsupervised driving performance assessment using free-positioned smartphones in vehicles," *in Proc. IEEE Int. Conf. Intelligent Transportation Systems Conf.*, Rio de Janeiro, Brazil, 2016, pp. 1598–1603.

[45] S. O. Sadjadi, A. Sathyanarayana, and J. H. L. Hansen, "Sensing variations in driving patterns using cost effective portable devices," presented at the Future Active Safety Technology Towards Zero Traffic Accidents Conf., Nagoya, Japan, Sept. 2013.

[46] A. Sathyanarayana, P. Boyraz, and J. H. L. Hansen, "Effects of multi-tasking on drivability through CAN-Bus analysis," in *Smart Mobile In-Vehicle Systems—Next Generation Advancements*, G. Schmidt, H. Abut, K. Takeda, and J. Hansen, Eds. New York: Springer, 2014, pp. 169–182.

[47] A. Sathyanarayana and J. H. L. Hansen, "Impact of secondary tasks on individual drivers: Not all drivers are created equally," *SAE Int. J. Passeng. Cars—Electron. Electr. Syst.*, vol. 5, no. 2, pp. 414–420, Sept. 2012

[48] N. Shokouhi, A. Sathyanarayana, O. Sadjadi, J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, Vancouver, Canada, May 2013, pp. 2834–2838.

[49] J.-C. Chien, J.-D. Lee, and L.-C. Liu, "A fuzzy rules-based driver assistance system," *Math. Probl. Eng.*, vol. 2015, pp. 1–14, Mar. 2015.

[50] L. Fletcher and A. Zelinsky, "Driver inattention detection based on eye gaze-road event correlation," *Int. J. Robot Res.*, vol. 28, no. 6, pp. 774–801, June 2009.

[51] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," *Real-Time Imaging*, vol. 8, no. 5, pp. 357–377, Oct. 2002.

[52] K. K. Ahlstrom and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, June 2013.

[53] A. Georgeon, M. J. Henning, T. Bellet, and A. Mille, "Creating cognitive models from activity analysis: a knowledge engineering approach to car driver modeling," in *Proc. Int. Conf. Cognitive Modeling (ICCM 2007)*, Ann Arbor, MI, July 2007, p. 26.

[54] E. C. B. Olsen, "Modeling slow lead vehicle lane changing," Ph.D. dissertation, Industrial and Systems Engineering, Virginia Polytechnic Inst. State Univ., Blacksburg, VA, 2003.

[55] K. A. Brookhuis and D. de Waard, "Assessment of drivers' workload: Performance and subjective and physiological indices," in *Stress, Workload, and Fatigue: Human Factors in Transportation*, P. Hancock and P. Desmond, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 321–333.

[56] M. Recarte and L. Nunes, "Mental workload while driving: Effects on visual search, discrimination, and decision making," *J. Exp. Psychol. Appl.*, vol. 9, no. 2, pp. 119–137, June 2003.

[57] Y. Liang and J. D. Lee, "Combining cognitive and visual distraction: Less than the sum of its parts," *Accident Anal. Prev.*, vol. 42, no. 3, pp. 881–890, May 2010.

[58] A. Azman, Q. Meng, and E. Edirisinghe, "Non-intrusive physiological measurement for driver cognitive distraction detection: Eye and mouth movements," in *Proc. Int. Conf. Advanced Computer Theory Engineering (ICACTE 2010)*, Chengdu, China, Aug. 2010, p. V3-595.

[59] M. Kutila, M. Jokela, G. Markkula, and M. R. Rue, "Driver distraction detection with a camera vision system," in *Proc. IEEE Int. Conf. Image Processing (ICIP 2007)*, San Antonio, TX, Sept. 2007, pp. 201–204.

[60] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 51–65, Feb. 2015.

[61] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 340–350, June 2007.

[62] M. C. Su, C. Y. Hsiung, and D. Y. Huang, "A simple approach to implementing a system for monitoring driver inattention," in *Proc. IEEE Int. Conf. Systems Man and Cybernetics (SMC-2006)*, Taipei, Taiwan, Oct. 2006, pp. 429–433.

[63] F. Tango and M. Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 894–905, June 2013.

[64] Q. Wu. "An overview of driving distraction measure methods," in *Proc. IEEE 10th Int. Conf. Computer-Aided Industrial Design and Conceptual Design (CAID CD 2009)*, Wenzhou, China, Nov. 2009, pp. 2391–2394.

[65] N. Li and C. Busso. "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE Int. Conf. Multimedia and Expo (ICME 2013)*, San Jose, CA, July 2013, pp. 1–6.

[66] N. Li and C. Busso, "Using perceptual evaluation to quantify cognitive and visual driver distractions," in *Smart Mobile In-Vehicle Systems—Next Generation Advancements*, G. Schmidt, H. Abut, K. Takeda, and J. H. L. Hansen, Eds. New York: Springer, 2014, pp. 183–207.

[67] N. Li, J. J. Jain, and C. Busso, "Modeling of driver behavior in real world scenarios using multiple noninvasive sensors," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1213–1225, Aug. 2013.

[68] J. Harbluk, Y. Noy, P. Trbovich, and M. Eizenman, "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance," *Accident Anal. Prev.*, vol. 39, no. 2, pp. 372–379, Mar. 2007.

[69] J. L. Harbluk, M. Eisenman, and Y. I. Noy, "The impact of cognitive distraction on driver visual behaviour and vehicle control," Transport Canada, Ottawa, Ontario, Tech. Rep. TP# 13889 E, 2002.

[70] M. H. Kutila, M. Jokela, T. Makinen, J. Viitanen, G. Markkula, and T. W. Victor, "Driver cognitive distraction detection: Feature estimation and implementation," *P. I. Mech. Eng. D-J Aut.*, vol. 221, no. 9, pp. 1027–1040, Sept. 2007.

[71] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.

[72] T. Misu, "Visual saliency and crowdsourcing-based priors for an in-car situated dialog system," in *Proc. Int. Conf. Multimodal Interaction (ICMI 2015)*, Seattle, WA, Nov. 2015, pp. 75–82.

[73] S. Jha and C. Busso, "Analysis of head pose as an indicator of drivers' visual attention," in *Proc. 7th Biennial Workshop Digital Signal Processing for In-Vehicle Systems and Safety*, Berkeley, CA, Oct. 2015, p. 15.

[74] H. Zhang, M. Smith, and R. Dufour. (2008). A final report of safety vehicles using adaptive interface technology (Phase II: Task 7C): Visual distraction. Delphi Electronics and Safety. Kokomo, IN. [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiRrp7li_PTAhVB5iYKHTMRCLUQFggiMAA&url=https%3A%2F%2Fwww.volpe.dot.gov%2Fsites%2Fvolpe.dot.gov%2Ffiles%2Fdocs%2FSAVE-IT%2520-%2520Visual%2520Distraction.doc&usg=AFQjCNEDi2l1SmiLCsSCqAOirWn8hOuWrw

[75] A. Tawari and M. M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *Proc. 2014 IEEE Intelligent Vehicles Symp.*, Dearborn, MI, June 2014, pp. 344–349.

[76] M.-C. Chuang, R. Bala, E. Bernal, P. Paul, A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Columbus, OH, 2014, pp. 165–170.

[77] W. H. Zangemeister and L. Stark, "Types of gaze movement: Variable interactions of eye and head movements," *Experimental Neurology*, vol. 77, no. 3, pp. 563–577, Sept. 1982.

[78] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *Proc. IEEE Int. Conf. Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, Nov. 2016, pp. 2157–2162.

[79] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," submitted for publication.

[80] C. Carter and R. Graham, "Experimental comparison of manual and voice controls for the operation of in-vehicle systems," *Ergonomics New Millennium*, vol. 44, no. 20, pp. 283–286, July 2000.

[81] J. Engelbrecht, M. J. Booysen, G. J. van Rooyen, and F. J. Bruwer, "Survey of smartphone-based sensing in vehicles for intelligent transportation system applications," *IET Intell. Transp. Syst.*, vol. 9, no. 10, pp. 924–935, Nov. 2015.

[82] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, Alcala de Henares, Spain, June 2012, pp. 234–239.

[83] M. V. Ly, S. Martin, and M. M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, Gold Coast, Australia, June 2013, pp. 1040–1045.

[84] P. Phondeenana, N. Noomwong, S. Chantranuwathana, and R. Thitipatanapong, "Driving maneuver detection system based on GPS data," presented at the Future Active Safety Technology Towards Zero Traffic Accidents Conf., Nagoya, Japan, Sept. 2013.

[85] Y. Zheng and J. H. L. Hansen, "Mobile-UTDrive: An android portable device platform for in-vehicle driving data collection and display," presented at the Future Active Safety Technology Towards Zero Traffic Accidents Conf., Gothenburg, Sweden, Sept. 2015.

[86] Y. Zheng, X. Shi, A. Sathyanarayana, N. Shokouhi, and J. H. L. Hansen, "In-vehicle speech recognition and tutorial keywords spotting for novice drivers' performance evaluation," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, June–July 2015, pp. 168–173.

[87] Y. Zheng, N. Shokouhi, A. Sathyanarayana, N. B. Thomsen, and J. H. L. Hansen, "Towards developing a distraction-reduced hands-off interactive driving experience using portable smart devices," presented at the SAE World Congr., Detroit, MI, Apr. 2016.

[88] Y. Zheng and J. H. L. Hansen, "Free-positioned smartphone sensing for vehicle dynamics estimation," presented at the SAE World Congr., Detroit, MI, April 2017.

[89] Y. Zheng, Y. Liu, and J. H. L. Hansen, "Navigation-orientated natural spoken language understanding for intelligent vehicle dialogue," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, Redondo Beach, CA, June 2017.

[90] Y. Zheng, Y. Liu, and J. H. L. Hansen, "Intent detection and context extraction for navigation dialogue language processing," submitted for publication.

[91] Transportation Research Board. [Online]. Available: http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx

SP

Craig T. Jin, Matthew E.P. Davies, and Patrizio Campisi

Embedded Systems Feel the Beat in New Orleans

Highlights from the IEEE Signal Processing Cup 2017 Student Competition

Foot-tapping and moving to music is such a natural human activity, one may assume that feeling the beat in music is a simple task. Feeling the beat and then producing it, e.g., by foot tapping, is an intrinsically real-time process. As listeners, we do not wait for the beat to occur before tapping our foot; instead, we make predictions about when the next beat in the music will occur and continually revise our sense of the beat based on the accuracy of our predictions. Likewise, performing musicians have shared sense of beat, which is what allows them to play in time together.

This type of high-level music listening and understanding sits at the heart of the challenge set for this year's IEEE Signal Processing Cup (SP Cup) competition, the final stage of which concluded at the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), hosted in New Orleans, America's jazz heartland, on 5 March. The participating undergraduate cohort had to devise and construct a creative, embedded application demonstrating a real-time response to the beat of the music. Depending on the genre, composition, and rhythmic complexity of a musical piece, real-time beat tracking poses considerable challenges, which are equally present for human listeners, especially those without formal musical training. Throughout the SP Cup, the teams confronted these challenges from both the human and computational perspectives via the choice of training and testing material, the human annotation of beat locations, the implementation and evaluation of their beat-tracking algorithms, and the response to the beat in their creative applications.

Digital Object Identifier 10.1109/MSP.2017.2698075
Date of publication: 11 July 2017

Beat tracking in music signals

The task of beat tracking of music signals has been an active area of music signal processing research for more than 25 years. While many of the earliest computational approaches sought to emulate the human process of tapping the beat in real time by making predictions of future beats [1], [2], a marked shift occurred in the early-to-mid 2000s toward offline approaches that could observe the entire musical input prior to determining beat locations.

The standard pipeline for offline beat tracking involves the explicit identification of note onset locations (or an "onset strength function," which emphasizes their location) that are subsequently passed to a tempo-estimation stage used to estimate the latent beat periodicity in the input signal, followed by the recovery of the phase (or alignment) of the beats to the music. Common techniques used to extract the beat from music signals include multiagent systems, dynamic programming, hidden Markov models, and a mixture of experts systems. Current state-of-the-art methods employ deep neural network architectures to learn the relationship between labeled beat annotations in training data sets and feature representations extracted from musical audio signals, thus leveraging both advanced signal processing and machine learning.

The growth of offline approaches arose in part by the significant increase in the use of beat tracking for so-called beat-synchronous analysis as an intermediate processing step within other music signal analysis tasks, such as structural segmentation, chord detection, and music transcription. With the shift toward making multiple passes across input signals, the focus on real-time analysis was reduced. Furthermore, with a greater emphasis on the accuracy of beat tracking over computational efficiency, offline approaches also provided the opportunity for tracking the beat in music with expressive timing (i.e., changes in tempo) something that was considered impossible for real-time systems bound by the need to make predictions of future beats in the music [3].

An emerging topic related to the domain of music signal processing is creative music information retrieval, which seeks to open new possibilities for music creation, interaction and manipulation. This is facilitated by the robust analysis and interpretation of music signals [4]. For applications that target live interaction between users and/or musicians and technology, there is a compelling need to perform music signal analysis in real time. One specific motivation for the SP

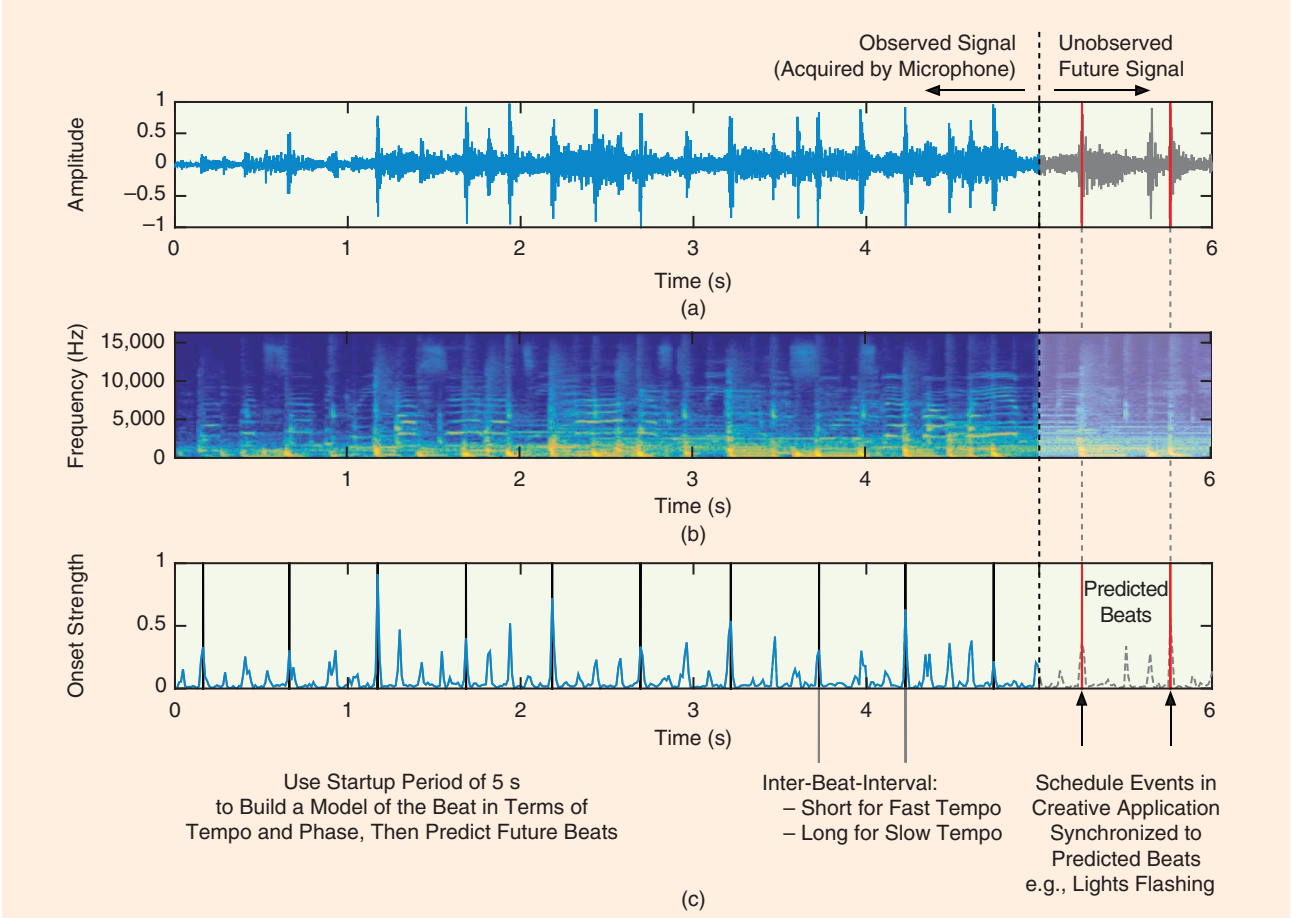


FIGURE 1. An overview of real-time beat tracking. (a) An input audio signal for which the first 5 s have been acquired by a microphone. (b) A spectrogram representation of the input audio signal used to generate the onset strength function. (c) The onset strength function with overlaid beat estimates shown in black and predicted future beats in red.

Cup was, therefore, to reimagine research into beat tracking with an explicit link to real-time creative applications. From a technical perspective, real-time beat tracking, unlike offline approaches, must extract an onset strength function, estimate tempo and predict future beats based only on a continuously evolving observation of the input signal, and thus it sits firmly at the more challenging end of the spectrum. This real-time requirement also imposes strict computational limitations, a difficulty that is only increased by constraining the use of hardware in the SP Cup to embedded devices with limited computational resources. The final aspect of the competition—developing a creative application that reacts to the (predicted) beat of the music—provides an open-ended activity for the teams, but one that must be also performed in real time on the embedded device. An overview of the process of real-time beat tracking is shown in Figure 1.

The SP Cup is an undergraduate competition organized by the IEEE Signal Processing Society (SPS) in which undergraduate students work in teams to tackle a real-life signal processing problem. Launched in 2014, the SP Cup competition has been held annually, and 2017 is the fourth edition.

To join, undergraduate students are required to form a team. Each team is composed of one faculty member to advise the team members, up to one graduate student to assist the supervisor in mentoring the team, and three to ten undergraduate students. Three top teams are selected from the initial round of competition and provided travel grants to participate in the final competition at 2017 ICASSP. The final results are described in "Winners of the SP Cup 2017."

Tasks in the SP Cup 2017

The SP Cup challenge covered the many and multidisciplinary aspects of beat tracking, with the aim of giving students training in several areas such as music understanding and beat annotation, strategies for selecting content for training and competition, signal processing, computational optimizations for real-time performance, hardware implementations, and creative application design and development. With such a wide range of tasks and challenges to address, the SP Cup 2017 was seen as the most challenging edition so far. All of the resources related to the competition can be found at http://sydney.edu.au/engineering/electrical/carlab/beatracking.htm.

The open competition stage

The SP Cup started with an open competition stage from June 2016 to January 2017, consisting of two parts. The

objective for part one was to submit three 30-s musical excerpts with human-annotated beat times. The judging criteria was the quality of the beat annotations. For the first part, participants were provided with a database of 50 musical excerpts spanning a range of styles and difficulties. The database was split into two halves. One half was open, meaning that for these musical excerpts, human-annotated beat times were provided. The other half was closed so that the annotated beat times for these musical excerpts remained hidden. The purpose of the database was to assist with the development and testing of real-time beat-tracking algorithms. The task for the first part consisted of an exercise in crowdsourced beat annotation. Each participating team was required to provide human-annotated beat times for three musical excerpts of their own choosing. They also nominated one of the three musical excerpts as a challenge piece so that the beat annotations would remain hidden from the other participating teams.

To assist participants with the evaluation of their beat-tracking algorithms and give a reference for how beat-tracking accuracy would be calculated, a MATLAB evaluation script was provided. The evaluation method, extended from [5], gives an accuracy score based on a comparison of estimated beat times with annotated ground truth. It calculates the proportion of continuously correct beat estimates occurring with a perceptually specified tolerance window around the ground truth annotations. To mirror the ambiguity in human perception of the beat in music, estimated beats at perceptually related metrical levels to the ground truth annotations (e.g., twice or half the tempo of the ground truth for music in 4/4 time) were also considered correct.

The first part of the open competition was devised to serve multiple purposes. From the perspective of the teams wishing to participate, the annotation of three musical excerpts provided a relatively low barrier for entry, while also offering teams the chance to actively shape the SP Cup through their personal choice of musical content. For the organizers, the use of team submitted content led to the creation of a totally new annotated data set for beat tracking (free from sampling bias) and, furthermore, one that could reflect the cultural diversity of the teams who participated.

For the second part of the open competition, participants had to develop and implement their beat-tracking algorithm on an embedded device (the choice was left open, but most used the Raspberry Pi for beat tracking and an Arduino for control of the output) so that it achieved real-time performance. The objectives for part two were the following:

1) real-time embedded software with instructions on how to run it
2) beat-time output for the real-time embedded device for the database and participant submitted musical excerpts
3) a video demonstrating real-time operation
4) a report in the form of an IEEE conference paper.

The judging criteria were a performance score for the real-time embedded algorithm and a creative application score. Participants then had to design and construct a creative application for their real-time beat-tracking device. In addition to submitting the beat-tracking output of their systems across all of the available musical material as well as providing source code with installation instructions, participants also had to submit a report in the form of an IEEE conference paper and post a video online demonstrating the creative application and real-time operation. This year's SP Cup is unique in that the competition included real-time constraints as well as a creative application.

The teams were evaluated on three main components submitted across both parts of the open competition. In the first part, a team of experts active in beat-tracking research assessed the subjective quality of the annotations and made corrections where necessary so as to ensure their validity as ground truth. In the second part, the submitted beat times provided by each team on the musical material without released annotations were evaluated using the publicly available MATLAB script. In addition, the creativity of the demonstrated applications were assessed, again by a group of experts. Since each team submitted the beat-tracking software for their real-time embedded device as part of the submission for the open competition, the real-time operation and its beat-tracking output could be verified. The final score for each team was weighted across these three components with the following proportions: one-sixth for the annotations, one-half for the real-time beat-tracking accuracy, and one-third for the creative application. A breakdown of the scores as well as a written assessment by the organizing committee was provided to all teams that participated in the second part of the open competition.

Final competition

After the judging committee evaluated the submissions from the open competition, three finalist teams were chosen to advance to the final competition. Prior to attending the final event at ICASSP, each team was required to submit additional annotated challenge excerpts to be used for on-site evaluation. However, in contrast to earlier stages in this year's competition, neither the audio nor the annotations were made available to the other teams.

The final SP Cup event was held at ICASSP in New Orleans, Louisiana, on 5 March. For the first time since the inception of the SP Cup, a live demo session was included in the final event. The event started by testing the accuracy of the real-time beat-tracking embedded devices in real-world conditions with the audio of the newly submitted challenge pieces captured by microphones (Figure 2). The finalist teams were then allowed time to set up their live demos. Each team then presented its beat-tracking algorithm, its implementation, and the design and development of the creative application. This was followed by a live demonstration of the creative application and a question and answer session. The final judging committee convened and selected the first-, second-, and third-prize winners as well as presented honorable mentions.

Winners of the SP Cup 2017

Grand Prize: Team Beats on the Barbie

- University of New South Wales
- Undergraduate students: Angus Keatinge, Max Fisher, Jeremy Bell, and James Wagner
- Supervisor: Vidhyasaharan Sethu
- Video: https://www.youtube.com/watch?v=VkoGZnVEsfw
- Technical Approach: Team Beats on the Barbie (Figure S1) adapted and optimized an existing real-time beat-tracking algorithm [6] for Rasberry Pi. They controlled their creative application, a robotic drumming system (see Figures S2 and S3), using an Arduino Mega. The robotic drummer can play back a drum part encoded as an Arduino sketch, and during the final competition it accompanied team members Jeremy Bell and James Wagner in a performance of John Lennon's "Imagine." Due to the use of high-powered solenoid drivers and fast triggers, the system was able to play drum fills and was loud enough to require no additional amplification.

Second Prize: Team Madmom

- Johannes Kepler University, Austria, and Télécom ParisTech, France
- Undergraduate students: Amaury Durand (Télécom ParisTech), Sebastian Pöll (Johannes Kepler University), and Raminta Balsyte (Johannes Kepler University)
- Supervisor: Sebastian Böck
- Graduate Mentor: Florian Krebs
- Video: https://www.youtube.com/watch?v=Losv4GqsGYU
- Technical Approach: Team Madmom (Figure S4) adapted a real-time beat-tracking system from the existing offline approach in the Madmom Python library [7] and used a recurrent neural network. To allow real-time operation, the bidirectional neural network was replaced with a unidirectional network. They controlled their creative application, a robotic drumming system (see Figures S5 and S3), using a Raspberry Pi. Instead of a preprogrammed drum pattern, the system inferred what to play based on the analysis of the rhythmic

FIGURE S1. First place team: Beats on the Barbie.

FIGURE S3. The automated drums for Team Madmom and Team Beats on the Barbie. Both teams implement drum signals for the bass drum, the snare drum, and the hi-hat.

FIGURE S2. Solenoid-based actuators for Team Beats on the Barbie.

FIGURE S4. Second place team: Team Madmom.

structure of the input and was able to react to changes in a time signature. Team Madmom intends to make its system freely available and open source at https://gitlab.cp.jku.at/ROBOD.

Third Prize: Team PulseBox

- University of Maryland, United States
- Undergraduate students: William Heimsoth, Creed Gallagher, and Josh Preuss
- Supervisor: William Hawkins
- Video: https://www.youtube.com/watch?v=KPwFnY6bJNI
- Technical Approach: Team Pulsebox (Figure S6) developed all aspects of their system entirely from scratch. Their beat-tracking algorithm made use of a novel comb-snapping technique to maintain high temporal accuracy of the predicted beats and used machine learning to optimize multiple relevant parameters including those related to tempo adjustment, windows, and the choice of frequency bands. Their creative system, the PulseBox, (shown in Figure S7) was a light-emitting diode (LED) cube containing 245 LEDs arranged in a 7x7 grid on each of the five visible faces of the cube. The LEDs were individually configurable with 24-bit color and were programmed to react to the beat of the music with rotating shapes and patterns.

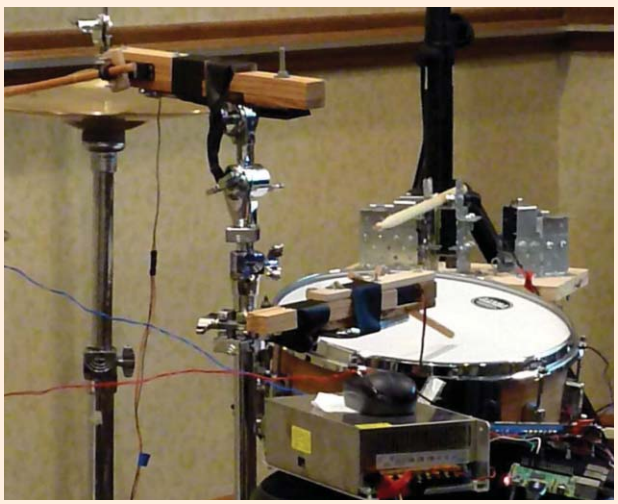


FIGURE S5. The drum actuators for Team Madmom.

In addition to the three overall winning teams (Figure S8), the SP Cup 2017 judging committee made the following honorable mentions. Videos for these and other submissions can be found at http://sydney.edu.au/engineering/electrical/carlab/beatracking.htm.

Honorable Mention for Excellent Video Production and an Entertaining Concept

- Team NTHU-EECS, National Tsing Hua University, Taiwan.

Honorable Mention for Excellent Video Production and Accurate Ground Truth Annotation

- Team Impulse, Bangladesh University of Engineering and Technology, Bangladesh.

Honorable Mention for Excellence in Ground Truth Annotation and Beat-Tracking Performance

- Team Sharif University of Technology, Sharif University of Technology, Iran.

FIGURE S6. The third place team: Team PulseBox.

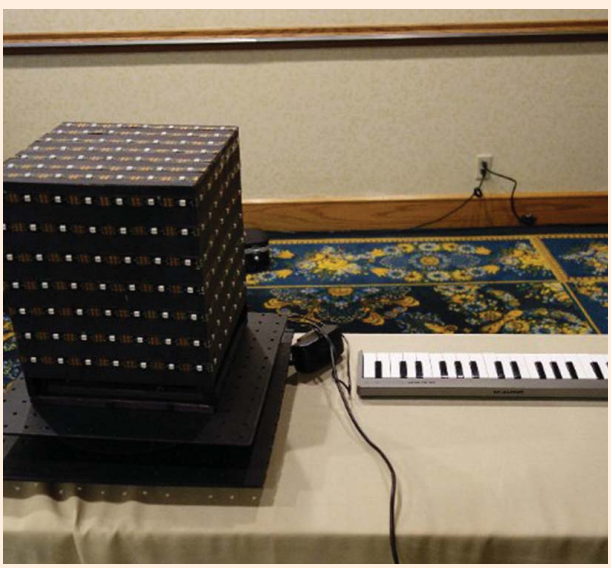


FIGURE S7. The rhythmic LED cube for Team Pulsebox.

Winners of the SP Cup 2017 (continued)

FIGURE S8. A behind-the-scenes look at the SP Cup 2017 teams that received honorable mentions: (a) Team Sharif, University of Technology, Iran; (b) and (c) Team NTHU-EECS, Taiwan, with their metronome mechanism; (d) and (e) Team Impulse, Bangladesh, and their band of skeletons.

Highlights of technical approaches

For the real-time beat-tracking aspect of the SP Cup, many teams implemented methods inspired by or directly adapted existing approaches for beat tracking. Even in the cases where a reference implementation was publicly available, these required a very significant overhaul to make the algorithms real-time compatible and sufficiently optimized to run on the embedded devices.

From an algorithmic perspective, the great majority of submitted algorithms followed the standard approach for beat tracking by

- generating one or more onset strength signals (often across subbands) derived from time-frequency representations of the streaming input audio signal
- performing periodicity analysis on the onset strength signal(s) by means of autocorrelation or comb filtering
- estimating the phase of the beats by cross-correlation or dynamic programming, and then using phase as the reference point from which to predict future beat locations.

Many teams also included some higher-level modeling to provide a smooth output without rapid switching between metrical levels (i.e., tempo doubling or halving). Depending on the computational resources of the chosen embedded device (some of which were extremely low power), the beat-tracking approach had to be highly optimized, e.g., purely based on time-domain analysis. The most computationally expensive and ambitious approaches attempted to run state- of-the-art deep neural network architectures for beat prediction.

The technical approaches were invariably biased by the initial project description, which mentioned blinking LEDs and the Raspberry Pi and Arduino. So, for example, the Raspberry Pi was the embedded platform used for beat tracking by the majority of teams (fourteen teams). A variety of other interesting embedded platforms were used by a single team: ARM mbed, NAO robot, STM32F4Discovery, and UDOO Quad. Many teams coupled the beat tracker with an Arduino to assist with the creative output. With regard to the programming language used for the embedded application, it was evenly distributed between C/C++ and Python. A wide variety of creative applications were demonstrated. Applications demonstrated by multiple teams were: LED displays (seven teams); screen displays (four teams); and automated drumming (two teams). The unique creative applications were a moving head, a dancing robot, a band of skeletons, a metronome follower, a vibration device for the hearing-impaired, and an encryption device.

SP Cup 2017 Statistics

In total, the teams from 20 different countries participated in SP Cup 2017. At the registration stage of the competition, 40 teams were involved with a total of 279 participants. For the first part of the open competition, 33 teams across 18 countries with more than 250 participants submitted musical excerpts (thus adding 99 new examples to the initial data set of 50 provided by the organizers). In the second part, which presented a significant increase in difficulty and submission requirements, 21 teams participated with 147 members spread across 14 countries. The countries with the most registrations were India with eight and the United States with seven.

As in previous years, the SP Cup was run as an online class on the Piazza platform, which, in addition to allowing continuous interaction with teams, also hosted the test material supporting documentation. In total, 115 students registered for the course, with approximately 220 contributions and 2,500 views of the posts. An archive of the class is available at https://piazza.com/ieee_sps/other/sp1701/home.

Since its inception, the SP Cup has received generous support from MathWorks, Inc., the maker of the popular MATLAB and Simulink platforms. MathWorks also provided funding support to the SP Cup and contributed their expertise. Each student team that registered for the SP Cup was provided complimentary software access to MATLAB and related toolboxes. After discussion with the SP Cup organizers, MathWorks provided skeleton code for real-time audio using a Raspberry Pi, which is available at http://au.mathworks.com/matlabcentral/fileexchange/59825-real-time-beat-tracking-templates-for-ieee-signal-processing-cup-2017. The IEEE SPS welcomes continued engagement and support from industry in future SP Cup competitions. Interested supporters may contact Dr. Patrizio Campisi, director for student services, at patrizio.campisi@uniroma3.it.

Participants' feedback

Throughout the open competition there was a great deal of interaction, not only through questions for the instructors posted to Piazza but also among the different student teams who often engaged in discussion over the provided responses. Indeed, these interactions were critical in expanding the flexibility of the evaluation script to correctly process music in non-4/4 meters. As organizers, we were delighted to see this collaborative spirit continue right through to the preparation for ICASSP and the final session itself. Next, we provide an overview of some feedback and perspectives received from the three winning teams.

FIGURE 2. A Real-time beat-tracking assessment for the final competition: music was played from the Bluetooth loudspeaker and recorded by three microphones, one for each team.

Team Beats on the Barbie

- "The project itself was extremely challenging. I worked on the software implementation of the algorithm, and to do this meant implementing the hardware interface on an embedded system. For me, the most challenging part of the SP Cup was setting up many different projects and libraries that often had never been tested on an embedded system to work in real time and simultaneously. This required running parts of the algorithm in different threads, modifying audio drivers, and writing low-level sound architecture code. Having these components running at the same time, and interacting with the hardware, was an amazing feeling."

  —Jeremy Bell, undergraduate

- "I learned a lot about DSP while working on the SP Cup, and since I am undertaking more DSP courses this semester, I feel more confident in my ability to understand more complicated concepts. I think my future career will almost certainly involve signal processing, so I will take the skills I have learned in DSP beyond university as well."

  —Jeremy Bell, undergraduate

- "I learned a lot about DSP algorithm design in general. I am also more confident in my understanding of sound architectures in Linux. I think I also learned a lot about teamwork, and what it takes to get things done under extreme time constraints."

  —Jeremy Bell, undergraduate

- "ICASSP was my first conference as an undergraduate, and I found it incredible. The amount of state-of-the-art technology and innovative creations was overwhelming, and it was almost impossible to keep up with in lectures. I was also surprised by the number of social events that occurred at the conference. It was great to be able to interact with so many talented and like-minded people on such a casual and friendly basis throughout the conference."

  —Jeremy Bell, undergraduate

- "We have already received several offers for other events at which we will be demonstrating the system. To do this will require some refinement of the interface and additional work on the software to make it more robust. Upon the graduation of our team, we will also be creating a handover document, so that future students can continue working on the system."

  —Team Beats on the Barbie

Team Madmom

- "I am interested in all topics making the link between music and mathematics, machine learning. I was working on incorporating online and real-time processing in the Madmom library when Sebastian told me that this work would be really useful for the SP Cup."

  —Amaury Durand, undergraduate

- "[Attending ICASSP was] really rewarding, it was my first time at a conference and, even though it was difficult for me to understand the talks I went to, I found it really interesting to meet the people who work on the topics that interest me."

  —Amaury Durand, undergraduate

- "[Participating in the SP Cup] was a perfect match. I just finished my Ph.D. in (mostly) offline beat and downbeat tracking, so it was very exciting for us to see how we can transform our system to work online and on an embedded device. Of course, it was more work than expected, but definitely a very exciting and rewarding experience!"

  —Florian Krebs, graduate mentor

- "The organization of everything was great, and I think there is no way to make this better. It was really great that you could organize a drum set, although this was not planned beforehand and not easy in a city that you don't know."

  —Florian Krebs, graduate mentor

- "It was very challenging given the limited processing power of the embedded device and extremely rewarding that it worked."

  —Sebastian Böck, supervisor.

Team Pulsebox

- "When I first heard of the topic for the 2017 SP Cup, I was very excited. As someone with a strong interest in both music theory and programming, I knew I had to get involved."

  —Creed Gallagher, undergraduate

- "One thing I learned a lot about was how to write truly speed-optimized code (Python with heavy use of NumPy). We had to push our Raspberry Pi to its limits. We also learned some lessons about the importance of effective communication and time management. We had to exercise a lot of discipline to complete such a big project on schedule."

  —Creed Gallagher, undergraduate

- "The signal processing challenge of beat tracking is incredibly complex! With so many types of songs and genres of music, there is no hard and fast rule as to what gives the best results. We ended up trying many approaches, many of which did not give as good results as we hoped. As a result, when we finally had something we felt performed well, it was incredibly satisfying."

  —Team Pulsebox

- "My senior project involves continual development of the PulseBox. I want to eventually create a 3-D holographic display that tracks both the beat and 'mood' of a song. Sebastian's team convinced us that the future of musical analysis lies in the use of neural networks, which is the avenue I will be exploring."

  —Creed Gallagher, undergraduate

- "As an undergraduate, attending ICASSP was an amazing and humbling experience. I enjoyed listening in on the presentations which gave me a window into the cutting edge of SP research. Plus, everyone was friendly and New Orleans was a fun venue."

  —Creed Gallagher, undergraduate

Forthcoming project competitions for undergraduates

The fifth edition of the SP Cup will be held at ICASSP 2018. The theme of the 2018 competition will be announced in September. Teams who are interested in the SP Cup competition may visit this link: https://signalprocessingsociety.org/get-involved/signal-processing-cup.

In addition to the SP Cup, the IEEE SPS recently announced the first edition of the Video and Image Processing Cup. The final competition will be held at the IEEE IEEE International Conference on Image Processing, in Beijing, China, 17–20 September. The theme of this competition is "Challenging Road Sign Detection." For details, visit: https://signalprocessingsociety.org/get-involved/video-image-processing-cup.

Acknowledgments

As the SP Cup 2017 Organizing Committee, we would like to express our gratitude to all of the people who made this adventure a reality: the participating teams, the judging panel, the local organizers, the IEEE SPS Membership Board for its financial support for the drum kit rental, and MathWorks for its sponsorship. Matthew E.P. Davies is supported by Portuguese National Funds through the FCT-Foundation for Science and Technology, I.P., under the project IF/01566/2015.

Authors

Craig T. Jin (craig.jin@sydney.edu.au) is an associate professor at the University of Sydney, Australia. He is a Senior Member of the IEEE and a current member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP TC). He initiated this edition of the SP Cup on behalf of the AASP TC, developed the competition specification and pedagogical materials, and ran the competition alongside Matthew Davies.

(continued on page 170)

Gianni Pasolini, Alessandro Bazzi, and Flavio Zabini

A Raspberry Pi-Based Platform for Signal Processing Education

One of the most important application areas of signal processing (SP) is, without a doubt, the software-defined radio (SDR) field [1]–[3]. Although their introduction dates back to the 1980s, SDRs are now becoming the dominant technology in radio communications, thanks to the dramatic development of SP-optimized programmable hardware, such as field-programmable gate arrays (FPGAs) and digital signal processors (DSPs). Today, the computational throughput of these devices is such that sophisticated SP tasks can be efficiently handled, so that both the baseband and intermediate frequency (IF) sections of current communication systems are usually implemented, according to the SDR paradigm, by the FPGA's reconfigurable circuitry (e.g., [4]–[6]), or by the software running on DSPs.

Introduction

The design of SDRs requires a broad expertise, which spans from high-level system architectures to SP techniques and programming languages, in addition to hardware-specific aspects [e.g., available memory, input/output (I/O) ports, timing issues, computational throughput]. This outspread know-how should be primarily acquired by communication-systems engineers during university courses, where the foundations of their expertise are laid. In our experience, however, providing students with a solid SDR background is a challenging goal, because of its twofold (theoretical and practical) nature. In fact, effective SDR teaching cannot disjoin theoretical fundamentals from practical laboratory experiments with real devices and instruments [7], [8]. Indeed, the connection between SDR theory and practice is so tight that it is desirable that both aspects are addressed within the same course, which would also allow the establishment of an active learning environment, with increased student involvement, motivation, and interest in the topics presented [9], [10].

Unfortunately, although laboratory facilities are usually available within engineering schools, practical SDR experiments are seldom proposed within courses, for two main reasons:

- the cost of the hardware (DSP and FPGA development kits), which can hardly be replicated in many working stations and, above all, left in the hands of inexperienced users
- the complexity and time requirements of such experiments, which should consist of several teaching phases concerning communication-system design, digital signal processing, DSP and FPGA programming, hardware-specific aspects, and system implementation on real hardware and measurements.

The latter issue is, in particular, very critical: usually, telecommunications and SP teachers are not interested in the specific syntax of the code used to implement a given subsystem (e.g., an IF modulator) or a given algorithm (e.g., a discrete Fourier transform), nor do they have the time for it within the limited duration of their courses.

In realtity, SP and telecommunications instructors focus on design methodologies, SP algorithms, as well as the analysis of signals in the time and frequency domains. On the other hand, they cannot assume that FPGA/DSP programming skills have been acquired by students during previous courses, as telecommunications/SP courses usually do not require such a background as a mandatory prerequisite. In many cases, therefore, practical SDR activities are either not provided to students or, at best, carried out using almost ready-made SDR implementations, thus weakening the beneficial "learning-by-doing" effect that is the primary goal of such activities, which gives the impression of incomplete teaching. The latter was, in particular, the case at the Engineering School of the University of Bologna, Italy, where SDR laboratory activities (that followed SP theoretical lectures) required students to write the missing parts of almost complete C-codes implementing SP algorithms on Texas Instruments' DSPs. The code skeleton provided to students was already complete in those parts concerning hardware dependent aspects (such as memory addresses, I/O management) that were not of interest from the SP teaching perspective.

Both of the aforementioned issues may be overcome, however, by the recent introduction in the mass market of general-purpose programmable devices with two fundamental characteristics:

- a limited cost (few-dozen dollars)
- the possibility to make the devices work with software automatically generated starting from the functional model (i.e., the block scheme) of the system to be implemented.

The Raspberry Pi board [11] is, perhaps, the most relevant example of such

Digital Object Identifier 10.1109/MSP.2017.2693500
Date of publication: 11 July 2017

devices: it is a popular, low-cost, single-board computer developed by the Raspberry Pi Foundation; its intention is to promote the teaching of basic computer science. It is largely used as a server in home networks or as a media center, and it is also adopted in many academic courses [12], mainly to teach computing and networking. What is particularly relevant for the SP educational purpose addressed in this article is that MathWorks provides a support package, which is an add-on software component for the use of third-party hardware, specifically addressing the connection between the Raspberry Pi and Simulink/MATLAB [13]. This so-called marriage allows the realization of a powerful, yet simple, SP educational platform that we have used for a couple of years to teach the basics of SDRs at the University of Bologna.

Simulink allows the graphical modeling and the simulation of the system to be implemented and then translates the model into software, which is finally downloaded on the device, thanks to the automatic code generation carried out by the support package. This allows students to design and implement on real hardware—even complex SP algorithms and SDRs—with no programming skills, thus saving the time usually required to become familiar with the programming language, investigate hardware low-level aspects (e.g., memory or registers addresses), develop the software, and debug.

Main contribution

In this article, we show an unconventional and innovative usage of Raspberry Pi boards for teaching purposes: deprived of usual I/O peripherals, such as a monitor, keyboard, and mouse, and equipped with the cables and connectors depicted in Figure 1, Raspberry Pi boards can be used as low-cost DSPs for the implementation of SP algorithms and, more generally, SDR-based communication systems [14].

Using this device jointly with Simulink, we realized a platform for the development of practical SDR activities provided to students: students facing the experimental activities, equipped with the Raspberry Pi as well as one personal computer (PC) hosting Simulink and the basic instrumentation of a telecommunication laboratory (signal generator, oscilloscope, and spectrum analyzer), experience the full process of an SP/SDR project, from the design and the simulation to the hardware implementation and the subsequent measurements with real instruments. All of this was done with no need to learn any programming language.

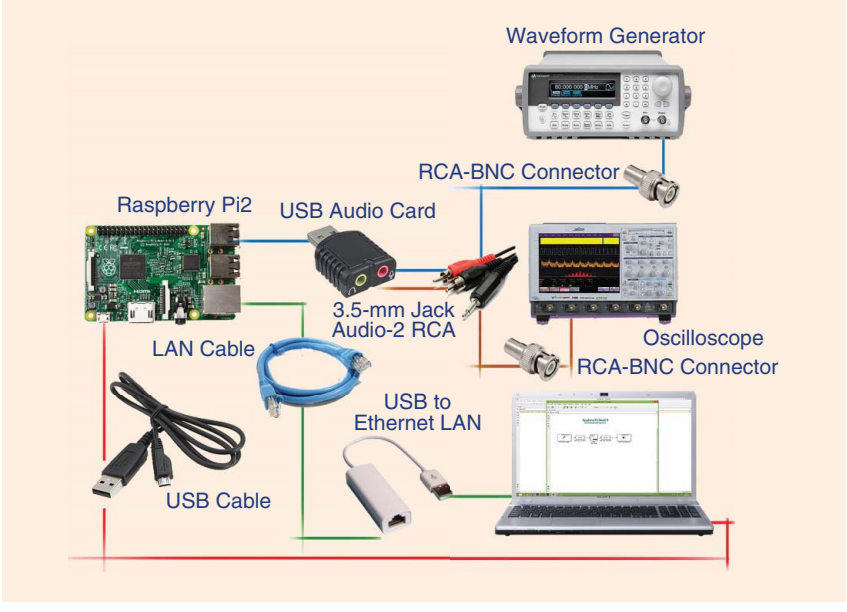


FIGURE 1. A block scheme of the workstation.

All of the material needed to implement the Raspberry Pi-based platform described in this article is freely available at www.simulinkdefinedradio.com [15]. The website provides

- a comprehensive document with the description of the platform setup, the hardware and software configurations, and the detailed description (system model, configuration, possible measurements) of the experimental activities developed so far
- the ready-to-use Simulink models of all systems that are the subjects of the experimental activities.

The hardware

Our SDR platform is based on the Raspberry Pi 2 Model B, the second-generation Raspberry Pi model shown in Figure 1; nevertheless, the SDR experiments we developed also can be carried out with the more recent Raspberry Pi 3 model. The Raspberry Pi 2 Model B is a credit-card-sized, single-board computer equipped with a quad-core Broadcom BCM2836 ARM v7 processor running at 900 MHz. Despite its low cost, at approximately US$40, it features 1 gigabyte of random-access memory (RAM), a 40-pin general-purpose I/O connector, four universal serial bus (USB) ports, a four-pole stereo output and composite video port, an HDMI port, a camera serial interfac connector, a display serial interface connector, a micro-secure digital (SD) card slot, and an Ethernet socket. It gained the attention of hobbyists and practitioners, especially for file server and media server applications. Today, there are plenty of existing projects, readily available on the Internet, which work on first-, second-, and third-generation Raspberry Pis.

In this article, we show an unconventional usage of this device, which is operated as a low-cost DSP for SP and SDR teaching. To achieve this goal, both an analog input and an analog output are required. Since the Raspberry Pi does not have natively an analog input, we used an external USB sound card with the 33051D chip set, like the one shown in Figure 1, that supplies a microphone input and an additional (with respect to the Raspberry Pi's) audio output.

This cheap device (aproximately US$20), connected to the Raspberry Pi's USB port, plays a double role: analog-to-digital converter (ADC) for input signals and digital-to-analog converter (DAC) for output signals.

Remark 1

As the sound card is conceived for audio signals, its sampling frequency is limited to 48,000 samples/s (with a resolution of 16 bits/sample). It follows that the band of signals that can be handled by our platform must be within the interval [0, 24] kHz. Apparently this is a severe limitation, as the carrier frequency of bandpass signals that we can manage must be in the order of 15~20 kHz and the bandwidth in the order of few kilohertz. However, the signal bandwidth is not a relevant issue for teaching purposes. In the case of digital transmissions, for instance, the bandwidth constraint simply entails that the bit rate must be kept properly low, but no modification of the system architecture is required. By way of example, the experimental activities we developed include the generation of digitally modulated signals with a bit rate up to 4.8 kilobits/second.

Furthermore, according to our experience, even when more sophisticated (and more expensive) platforms (e.g., Texas Instruments' C6748) are adopted for experimental activities within university courses, their audio ports are still usually adopted as analog I/O interfaces, with the same bandwidth limitation.

The software

Simulink, developed by MathWorks, is a graphical extension to MATLAB for the modeling and simulation of linear and nonlinear dynamic systems. In Simulink, systems are drawn on screen by means of simple drag-and-drop operations of elementary blocks, which are interconnected each other to realize the final block diagram. An example of such a diagram, denoted *model* in this article, is given in Figure 2, which shows the implementation of an orthogonal frequency-division multiplexing (OFDM) transmitter. Starting from the bit sequence at the output of the Bernoulli binary generator block (leftmost block), all steps of the OFDM

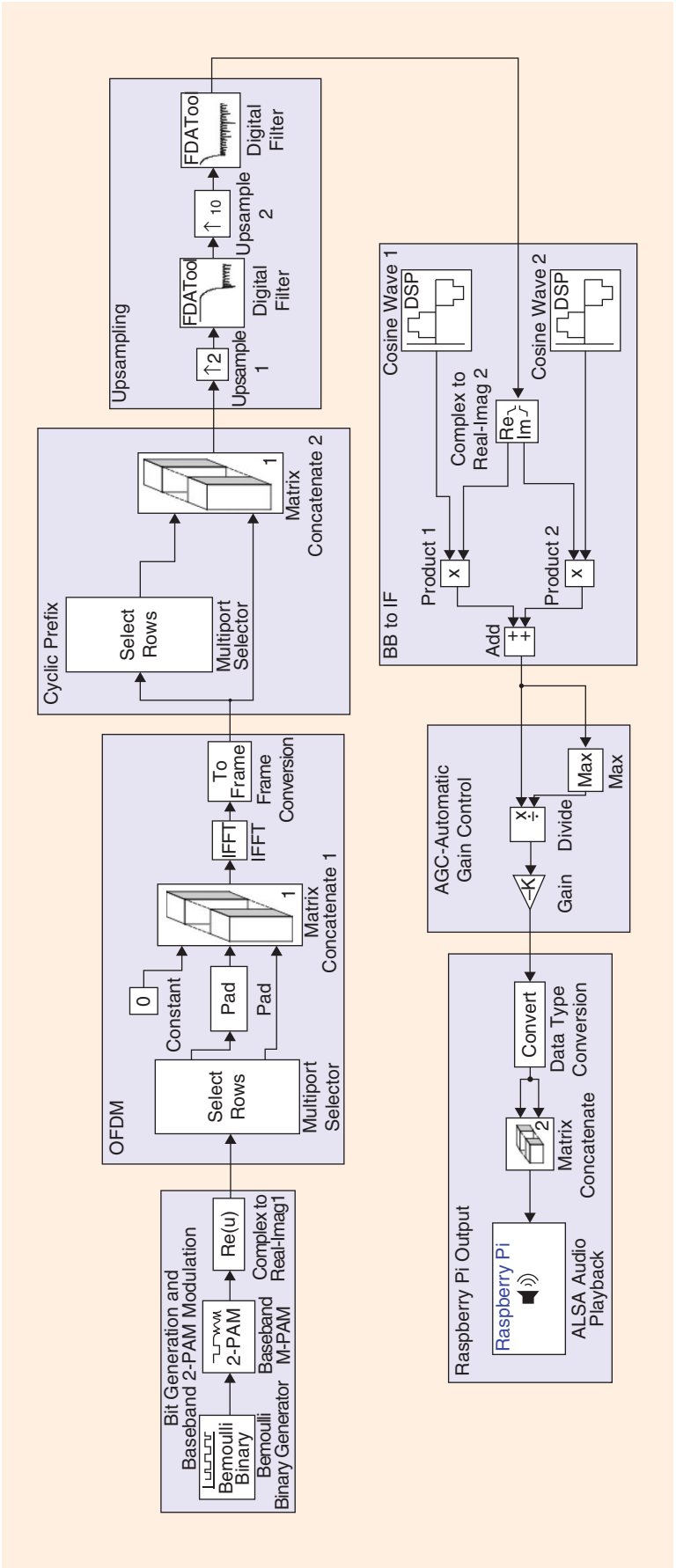


FIGURE 2. The Simulink model of the OFDM transmitter. AGC: automatic gain control. BB: baseband.

modulation are performed, which generate the output signal at the Raspberry Pi's audio port (represented by the block with the loudspeaker icon).

The elementary blocks are collected in a comprehensive library of toolboxes, that includes also virtual I/O devices such as function generators and oscilloscopes. The behavior of each elementary block is easily controlled by opening the corresponding configuration window and properly setting its parameters.

Once the model has been completed, its functioning can be checked starting the Simulink simulation, which benefits from the possibility of visualizing signals' characteristics (shape, spectrum, constellation, etc.) in each section of interest by means of the virtual I/O devices. Thanks to the free Raspberry Pi support package provided by MathWorks, Simulink is then capable of translating the designed model into low-level code, which is finally downloaded on the board and executed. The Raspberry Pi can now work as a standalone (it may be even unplugged from the PC), and can be connected to real instruments, thus allowing the system testing on real hardware with real signals.

From the perspective of the SP teaching methodology, this possibility is, without a doubt, a revolution. In fact, it permits the hardware implementation of complex systems by simply drawing and configuring the corresponding block diagram with Simulink, thus allowing

- an immediate visual correspondence with the functional block diagrams sketched on the blackboard during theoretical lectures
- the possibility to observe and measure the actual signals generated/received/processed by the system implemented
- removing the need to teach DSP/FPGA programming languages, focusing the attention on SP and systemistic aspects
- the realization cost of an SDR laboratory to be reduced.

Indeed, the automatic code generation functionality provided by Simulink supports several programmable devices specifically designed for SDR implementations, hence, much more performing than the Raspberry Pi. However, even putting aside the higher costs, such devices (e.g., Xilinx FPGAs and Texas Instruments' DSPs) usually require the installation of additional (with respect to MATLAB/Simulink) software (e.g., System Generator for Xilinx or Code Composer Studio for Texas Instruments), which also entails possible license issues. In addition, working with such additional software usually requires administrator privileges on the PC, which are not granted to students in a university laboratory. To work around this, for example, in our previous experience at the University of Bologna, we were forced to install virtual machines on laboratory computers to avoid permission problems, which made the whole setup much more complicated than the one described in this article. This installation meant we did not require additional software (other than MATLAB/Simulink), which dramatically eases the workstation setup and usage.

The resulting SDR platform

Next, we will describe some didactic experiments concerning the Simulink modeling and the subsequent hardware implementation of telecommunication systems and digital SP algorithms. Apart from the Raspberry Pi board, such activities require a PC hosting MATLAB and Simulink, as well as instruments for the generation and analysis of signals in the frequency and time domains. The equipment includes also cables and adapters suitable for interconnecting the Raspberry Pi to both the PC and the instruments. The whole setup is detailed next.

The software setup

First, the platform requires a PC equipped with MATLAB and Simulink. In particular, the Simulink models of the implemented systems have been realized with MATLAB R2015a equipped with the following libraries:

- Signal Processing Toolbox
- DSP System Toolbox
- Communications System Toolbox
- Raspberry Pi Support Package.

The three toolboxes, which require a valid MATLAB license, provide the elementary blocks to generate, process, and visualize signals in the time and frequency domains. They comprise all of the algorithms needed to compose the physical layer model of a communication system, with particular reference to data generation, channel coding, modulation, filtering, demodulation, and carrier- and symbol-timing synchronization, also including the tools to design and analyze finite impulse response and infinite impulse response filters, even in the multirate, multistage, or adaptive cases. They also provide virtual instruments to visualize constellations, eye diagrams, and power spectra, and obtain performance metrics such as bit error rate and error vector magnitude.

The support package, instead, can be downloaded at no charge and provides Simulink with an additional toolbox, which includes Raspberry Pi's specific blocks, such as those needed to drive its digital I/O and read and write data from them. Moreover, it updates the Raspberry Pi's operating system, which resides on a conventional micro-SD memory card, adding those features needed to establish a connection between the board and Simulink.

The hardware setup

The hardware setup is shown in Figure 1. It consists of one PC hosting Simulink, connected to the Raspberry Pi's Ethernet port by means of a USB-Ethernet adapter (or directly to an Ethernet port if there is one available). The Raspberry Pi's analog output, provided by the external USB sound card, is fed to an oscilloscope and/or a spectrum analyzer for signal analysis. Some experiments also require a function generator, whose task is to provide the Raspberry Pi with an external signal (such as the carrier when implementing a modulator). In such a case, the sound card's input port is used as well.

Observe that, although the Raspberry Pi is natively equipped with an integrated audio output (headphone output), it is surely preferable to use the analog output provided by the external sound card. The integrated DAC is, in fact, a low cost component with poor quality: although the distortion introduced on the output signal is almost inappreciable to a human ear, it appears well evident if the signal is observed with an oscilloscope (more details about this effect can be found in the documentation available in [15]). This

unexpected issue is solved using the additional good quality audio output provided by the external sound card.

The USB cable connecting the PC to the Raspberry Pi (Figure 1) is needed only to provide the latter with the required 5V dc power. Although a dedicated power supply (like those used for cell phones) could be used for the same purpose, we observed that this solution can cause some problems due to the disturbances on the output signal generated by the ac/dc converter. The occurrence and entity of this phenomenon depends on the quality of the converter. Powering the Raspberry Pi by means of the PC's USB port is the easiest way to bypass such an issue.

The full list of the SDR platform's components is reported in Table 1. The total cost of the hardware (excluding the PC) is approximately US$100.

Laboratory equipment

The experimental activities realized by means of the SDR platform discussed here result in the realization of systems able, in general, to generate and process signals.

The modeling stage, carried out using Simulink, and the consequent implementation on the Raspberry Pi are thus followed by a signal measurement phase, aimed at verifying through experimental observations the correct functioning of the designed system and the related theoretical concepts.

The equipment required for the measurement campaign is typically available in every didactic laboratory for electronics and telecommunications, with particular reference to an oscilloscope, a spectrum analyzer, and a signal generator. The signal generator, in particular, is mostly used as a sine wave generator, to provide the carrier needed by some of the implemented transmitters or the input signal for digital filtering systems.

The signals that will be generated/processed by our systems are within the [0 24] kHz band, owing to the characteristics of the external sound card introduced in the section "The Hardware." For the generation or analysis of such signals there is no need for sophisticated instruments; the basic instruments typically available in a didactic laboratory, generally solid but not highly performing, are suitable for the experimental activities described next.

Nonetheless, even basic oscilloscopes and spectrum analyzers are costly devices for students or hobbyists. Therefore, the experiments presented here apparently cannot be performed "at home," where laboratory instruments are usually not available. Indeed, this is not true. Recently, low-cost, multipurpose instruments have been conceived for low-frequency applications. In this regard, the Digilent Analog Discovery device [16] is worth special attention. When connected to the PC through the USB port, this device is able to generate and acquire signals. With a moderate price, in the order of US$270, this tool can operate as signal generator, oscilloscope, spectrum analyzer, network analyzer, logic analyzer, digital signal generator, and power supply. The user interface of each single instrument, displayed on the PC's monitor, shows the same knobs, sliders, and buttons of the full hardware instrument, allowing the user to perform the measurement activity as if he or she were in a laboratory. Of course, the bandwidth that can be handled by the Digilent Analog Discovery, in the order of tenth of megahertz, cannot be compared with that of more sophisticated and expensive instruments, however, it is more than adequate for the didactic experiments carried out with our SDR platform. In this case, therefore, the signal generator, the oscilloscope, and the spectrum analyzer can be conveniently replaced by this multifunction tool.

Developed SDR experiments

Several SP and SDR experiments have been developed at the University of Bologna using the previously described platform. In particular, the following systems have been already implemented and are proposed since a couple of years to students enrolled in the master's degree in telecommunications engineering:

- signal generation
- digital filtering
- adaptive noise canceler
- two-level pulse amplitude modulation (2-PAM) and 4-PAM baseband transmitters and receivers
- two-level amplitude shift keying (2-ASK) and 4-ASK transmitters and receivers
- frequency-shift keying transmitter
- quadrature phase-shift keying transmitter and receiver
- OFDM transmitter (64 and 256 subcarriers).

The corresponding Simulink models and the related documentation, conceived as material to be provided to students, are available for free download in [15].

According to our lecture organization, each experiment consists of a theoretical introduction on the system to be implemented, which provides the fundamentals

Table 1. The SDR platform components.

Material and quantity	Usage
Number 1 PC hosting Simulink	It is used for system design and automatic code generation.
Number 1 Raspberry Pi 2 (or 3)	It is used as a DSP.
Number 1 External sound card	It provides analog input and output ports.
Number 1 Micro-SD memory card	It contains the Raspberry Pi operating system.
Number 1 USB-micro USB cable	It is used to connect the Raspberry Pi to the PC's USB port for the power supply.
Number 1 USB to LAN adapter	It allows the connection to the Raspberry Pi to the PC using the USB port.
Number 1 Ethernet cable	It is used to connect the Raspberry Pi to the PC.
Number 2 3.5-mm-RCA jack cables	They are used to connect the sound card, equipped with 3.5-mm female jacks, to the instruments.
Number 2 RCA-BNC adapter	They are used to adapt the 3.5-mm-RCA connectors to the Bayonet Neill–Concelman (BNC) connectors of the instruments.

on its architecture, functioning, and performance, followed by the laboratory experiment. During the activities, students, assisted by tutors, are requested to develop the system model with Simulink, execute the corresponding (automatically generated) code on the Raspberry Pi, and carry out frequency-domain and time-domain analysis of the output signal with a spectrum analyzer and an oscilloscope. Developing the system model, designed from the outset with the hardware implementation objective, students get familiar with fundamental SP operations, such as digital filtering and multirate processing. They also face a number of issues arising from hardware constraints, such as the finite precision representation of numbers and the DAC/ADC dynamic ranges (thus facing possible quantization noise and nonlinear distortion due to saturation). They can even change the system parameters run time and observe the corresponding effect on the output signal.

Eventually, students get a complete picture of the system investigated, merging in an overall view the different perspectives provided by its high-level architecture, the constituent SP algorithms, the hardware characteristics, and the signals' measurements outcomes—all this with low-cost hardware that can be replicated in many working stations, with no need to learn any programming language.

Due to space constraints, only a couple of experimental activities are described next, which concern the implementation of a baseband 2-PAM transmitter and a passband OFDM transmitter. The interested reader is referred to the material provided in [15] for more information.

Examples of experimental activities

The experimental activities introduced in the section "Developed SDR Experiments" are proposed to students following an increasing complexity order. They start from the implementation of a simple signal generator, which produces at the Raspberry Pi's output a sinusoid with controllable frequency and amplitude, and conclude their laboratory activities with the design and experimental characterization of an OFDM transmitter with 64 subcarriers. In between, they implement the most typical baseband and passband digital transmitters and receivers, getting familiar with the basics of communication systems and the related SP techniques.

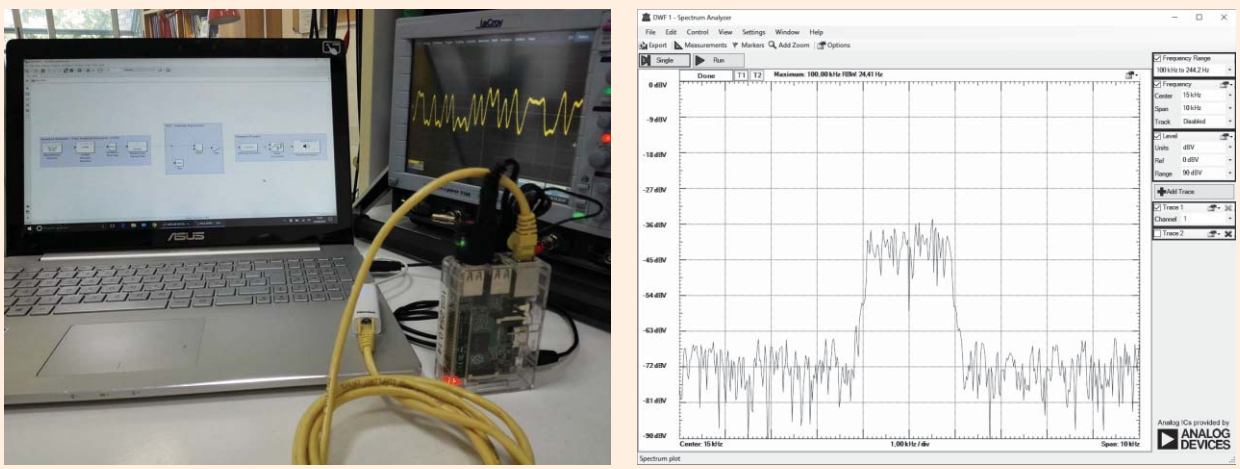
Example 1 (2-PAM transmitter)

Figure 3(a) shows, for instance, the whole platform (PC with Simulink and Raspberry Pi with external USB sound card) and an example of time-domain analysis in the case of a baseband 2-PAM transmitter with raised cosine pulse shaping. The system model, which is visible on the screen of the PC, is separated into three parts. In the first part, random bits are generated, converted into symbols of the $\{+1, -1\}$ alphabet, and passed through a shaping filter that generates a sampled raised cosine baseband signal. The following two parts convert the samples sequence into the analogue output signal and are common with most of the other transmitters. More specifically, in the second part the amplitude of the discrete-time signal is adjusted to fully exploit the DAC dynamic range whereas in the last part it is passed to the sound card's DAC, which generates the analog output. The model, translated by Simulink into the corresponding code and executed on the Raspberry Pi, generates the 2-PAM signal displayed by the oscilloscope.

Apart from putting into practice their knowledge on communication systems' design and SP techniques, students can observe in real time the impact of system's parameters (such as the bit rate) or signal's properties (such as the pulse shape) on the classic plots that characterize a digital signal, such as the eye diagram or the power spectrum.

Example 2 (OFDM transmitter)

In the last (and most complex) experiment, students are required to implement an OFDM transmitter with 64 subcarriers, whose model is shown in Figure 2. In this case, the block diagram has been divided in seven parts, with the last two being the same as those used in the previous example. In the first part, starting from the bits produced by the



(a) (b)

FIGURE 3. Examples of experimental activities: (a) 2-PAM baseband transmitter and (b) OFDM signal spectrum measured with a spectrum analyzer.

Bernoulli binary generator block at a data rate of 1.8 kilobits/s, the baseband M-PAM block generates the modulation symbols out of a $\{+1, -1\}$ alphabet, which modulate the 48 subcarriers used for pilot signals and data. Then, in the second part, the matrix concatenate1 block provides null data (the constant 0) to the central subcarrier, which corresponds to the dc subcarrier, and to the lateral subcarriers (fueled by the zero padding), which correspond to the virtual subcarriers. The 64 modulation symbols are thus arranged as required by the following inverse fast Fourier transform (IFFT) block. The baseband complex signal generated by the IFFT block, collected into frames, is then added of the cyclic prefix in the third part of the block scheme, upsampled by a two-stage interpolator in the fourth part, and upconverted to the IF band centered at 15 kHz in the fifth part. Proper amplification (sixth part) and digital-to-analog conversion (seventh part) complete the model.

This experiment, which implements a scaled version of an actual Wi-Fi transmitter (with lower carrier frequency and data rate), summarizes some issues already faced by students in previous experiments, such as multirate processing (see the upsampling macro-block in Figure 2) and signal dynamic range adjustment (see the automatic gain control needed to fit the DAC's dynamic range in Figure 2), with new issues, which are OFDM specific, such as data multiplexing, IFFT-based modulation, and cyclic prefix generation.

Connecting the platform's output to the spectrum analyzer, the signal spectrum appears as shown in Figure 3(b): the expected bandwidth of approximately 2 kHz and the dc subcarrier centered at 15 kHz clearly can be observed.

Challenges and lessons learned

All experiments find their conclusion in a verification phase, in which students take measurements with the oscilloscope or the spectrum analyzer and check the matching between what they observe and their expectation.

From the instructor's perspective, this step turned out to be more delicate than expected. On the one hand, in many cases students were not acquainted with the instruments and their settings. This required the introduction of a short tutorial on the basics of electronic measurements. On the other hand, this aspect also proved to be an excellent learning opportunity, as students were encouraged to explain their unexpected observations, which further enforced the connection between theory and practice.

Another lesson learned concerns the activity organization: individual work could be very fruitful, as each student is fully engaged in the experiments. However, this entails a significant effort on the instructor's side, who could be in the position to handle several simultaneous requests for assistance. In this case, the availability of skilled tutors is surely advisable. Teamwork can be an effective method to relieve such issue, as in many cases students might overcome possible difficulties through cooperative thinking and decision making, engaging with each other in thoughtful learning. Moreover, teamwork is the norm in almost any work environment, hence, it should be encouraged, when possible, also at the university level. Nonetheless, there is the risk that some students are less involved in the experiment than the other team components, which could reduce the effectiveness of the laboratory activity. It is the instructors' and tutors' responsibility to ensure that all students are actively committed.

Conclusions

After a couple of years of hands-on laboratory activities with the previously introduced SDR platform, some conclusions can be drawn, both from the perspectives of the teachers and the students. From our point of view, the immediate practical application of the theoretical knowledge acquired during classical lectures significantly accelerated the students' learning curve. In a handful of hours they pass from the passive learning of the teacher-centered lecture to the firsthand design and measurement of a real system, which broaden and strengthen their knowledge. Of course, this acceleration is a direct consequence of the automatic code generation capability provided by Simulink jointly with the simplicity (and affordability) of the platform setup.

As far as the students' experience is concerned, we encountered a very positive attitude. Working with real devices gives them new motivation, so that they generally approach the experimental activities with fresh enthusiasm. The low cost of the Raspberry Pi appealed to them and also encouraged some students to buy his or her own device to further develop SP experiments, which can be taken as encouragement to carry on our effort on the proposed platform.

Given the positive feedback, we are, in the near future, also considering including in our platform RF transmitters and receivers, which require the adoption of additional hardware for the RF part, such as the HackRF One [17] transmitter/receiver and the RTL-SDR receiver [18].

Acknowledgments

This activity has been carried out at the Wireless Communications Laboratory (WiLAB) of the University of Bologna/National Research Council. We thank WiLAB Director Prof. Oreste Andrisano for his comments and suggestions and for providing all technical facilities. We are also very grateful to Mirko Mirabella for his great contribution to the Simulink Defined Radio project. Finally, we are indebted to Stefano Olivieri (MathWorks) for his great willingness to discuss the details for our activity, providing us valuable suggestions. This work has been partly funded by MathWorks in the framework of the Academic Support A#: 1-2073488447.

Authors

Gianni Pasolini (gianni.pasolini@unibo.it) received his M.S. degree in telecommunications engineering and his Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 1999 and 2003, respectively. In October 2006, he became a researcher at the University of Bologna, where he has been teaching telecommunications since 2003. Now he is with the Wireless Communication Laboratory (WiLAB) of the same university, where he is working on digital signal

processing, wireless networks, and intelligent transportation systems. He serves as a reviewer for many transactions/journals and conferences and as a technical program committee member of several international conferences.

Alessandro Bazzi (alessandro.bazzi@ieiit.cnr.it) received his Laurea degree (with honors) and his Ph.D. degree in telecommunications engineering both from the University of Bologna, Italy, in 2002 and 2006, respectively. Since 2002, he has worked for the Institute of Electronics, Computer, and Telecommunication Engineering of the National Research Council of Italy, and, since the academic year 2006–2007, he has been acting as adjunct professor at the University of Bologna.

Flavio Zabini (flavio.zabini2@unibo.it) received his Laurea degree (summa cum laude) in telecommunications engineering and his Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 2004 and 2010, respectively. He is a researcher at the Wireless Communication Laboratory of the University of Bologna. In 2013–2014, he was a postdoctoral fellow at the German Aerospace Center, Cologne, Germany. His current research interests include channel coding for deep-space, echo cancelation, multidimensional random sampling, and performance-fairness tradeoff in communication systems. He serves as an editor of *KSII Transactions on Internet and Information Systems.*

References

[1] J. Mitola and Z. Zvonar. (2001). *Software Defined Radio Applications and Economics*. Wiley-IEEE Press, Piscataway, NJ, pp. 419–473 [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5263802

[2] T. Ulversoy, "Software defined radio: Challenges and opportunities," *IEEE Commun. Surv. Tut.*, vol. 12, no. 4, pp. 531–550, 2010.

[3] A. A. Abidi, "The path to the software-defined radio receiver," *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 954–966, May 2007.

[4] G. Chiurco, M. Mazzotti, F. Zabini, D. Dardari, and O. Andrisano, "FPGA design and performance evaluation of a pulse-based echo canceller for DVB-T/H," *IEEE Trans. Broadcast.*, vol. 58, no. 4, pp. 660–668, Dec. 2012.

[5] G. Pasolini and R. Soloperto, "Multistage decimators with minimum group delay," in *Proc. IEEE Int. Conf. Communications*, May 2010, pp. 1–6.

[6] F. Zabini, G. Pasolini, and O. Andrisano, "Design criteria for FIR-based echo cancellers," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 562–578, Sept. 2016.

[7] M. Petrova, A. Achtzehn, and P. Mhnen, "System-oriented communications engineering curriculum: Teaching design concepts with SDR platforms," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 202–209, May 2014.

[8] M. Simoni and M. Aburdene, "Lessons learned from implementing application-oriented hands-on activities for continuous-time signal processing courses," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 84–89, July 2016.

[9] H. I. Modell, "Preparing students to participate in an active learning environment," *Adv. Physiol. Educ.*, vol. 15, no. 1, pp. 6977, June 1996.

[10] H. L. Lujan and S. E. DiCarlo. (2006). Too much teaching, not enough learning: What is the solution? *Adv. Physiol. Educ.* [Online]. *30(1)*, pp. 17–22. Available: http://advan.physiology.org/content/30/1/17

[11] Raspberry Pi. (2017). [Online]. Available: http://www.raspberrypi.org

[12] Raspberry Pi Academy. (2017). [Online]. Available: http://www.raspberrypi.org/picademy/

[13] Raspberry Pi Support from Mathworks. (2017). [Online]. Available: http://it.mathworks.com/hardware-support/raspberry-pi-simulink.html

[14] G. Pasolini, F. Zabini, A. Bazzi, and S. Olivieri, "A software defined radio platform with Raspberry Pi and Simulink," in *Proc. IEEE 24th European Signal Processing Conf.* Aug. 2016, pp. 398–402.

[15] Raspberry Pi based SDR experiences. (2017). [Online]. Available: http://www.simulinkdefinedradio.com/

[16] Digilent analog discovery. (2017). [Online]. Available: http://store.digilentinc.com/analog-discovery-2-100msps-usb-oscilloscope-logic-analyzer-and-variable-power-supply/

[17] Hackrf One. (2017). [Online]. Available: https://greatscottgadgets.com/hackrf/

[18] RTL-SDR. (2017). [Online]. Available: http://www.nooelec.com/store/sdr/sdr-receivers/nesdr-mini-rtl2832-r820t.html

Waheed U. Bajwa

On "Flipping" a Large Signal Processing Class

Modern academy traces its roots back to the medieval universities established between the 12th and the 14th centuries [1]. Much has changed in the world of academia during the millennium that separates a modern university from a medieval one.

Among these changes, there are two that arguably stand out the most. First, university education is no longer considered the exclusive purview of a select few; rather, it has become a basic human right for all. Second, technology has become an integral component of university education, be it the delivery of information through multimedia presentations, the use of e-mail for student–teacher interactions, the reliance on course management systems for submission and grading of assignments, or the adoption of e-books as class texts. But there is one thing in academia that has remained largely unchanged since the advent of medieval university: the mode of instruction.

Lecturing—in which an instructor imparts knowledge to students by standing in front of them and reciting relevant information that is recorded by the attendees—was the only mode of instruction in medieval universities [1]. Lecturing remains the dominant mode of instruction in modern academy. This is despite the fact that research on learning indicates lecturing is not the most effective means of helping students master the course material [2]–[5]. The reason for the survival of lectures in modern academy is simple: among all

Digital Object Identifier 10.1109/MSP.2017.2698074
Date of publication: 11 July 2017

the modes of instruction available to today's instructors, lecturing remains the quickest and cheapest means of educating large numbers of students.

The purpose of this article is to argue, however, that a carefully "flipped" classroom can be used to replace a traditional lecture-based classroom with minimal time, cost, and infrastructure overhead—even for large classes with hundreds of students. The findings reported in this article are mostly based on my seminal 15-week flipped offering of a junior-level signal processing class. The final enrollment in the class was 133 students in the Department of Electrical and Computer Engineering at Rutgers, The State University of New Jersey, in the Spring 2016 semester.

The case against the lecture format

Tens of millions of students graduate from universities around the world in which instructions are centered around lectures. This is proof enough that lecturing works. Recent research, however, makes it abundantly clear that lecturing does not result in the best learning outcomes for all students [2]–[5]. This is perhaps more true in science, technology, engineering, and mathematics disciplines than in other disciplines. In particular, the following limitations of the lecture format in engineering education started me on my quest to seek more effective, but low-overhead, alternatives to lecturing.

The fallacy of academic equivalence

Engineering instructors all over the world will have no hesitation accepting that "no two students are alike academically." This truism holds regardless of whether one is an instructor at a more- or a less-selective university and whether one teaches a mandatory introductory course or an advanced elective class. The initial academic variation among newly admitted students can be primarily attributed to their diverse educational, geographic, and socioeconomic backgrounds. Afterward, the unavoidable pyramid structure of engineering curriculum begins to amplify this initial variation. However, the lecture format ignores the academic variation among students and, instead, makes the fallacious assumption that all students enrolled in a class have required mastery of prerequisite concepts. The unfortunate outcome of this "fallacy of academic equivalence" is that two students, one of whom secured an "A" and one of whom managed a "D" in the prerequisite course—receive identical instructions in the classroom.

The fallacy of behavioral equivalence

The lecture format is primarily a passive mode of instruction [13], with active interactions between the instructor and the students mainly taking place in two scenarios: 1) the instructor probes and/or prompts the students to gauge their understanding of the presented material, and 2) the students ask clarifying questions by interrupting the instructor. An instructor who relies on the lecture format for achieving the learning objectives of the class effectively makes an implicit assumption that students are capable of utilizing the aforementioned avenues for turning a passive lecture into an active one. Unfortunately, this is another fallacious assumption; just like academically, no two students are behaviorally alike! Indeed, for every student in a classroom who is apt at interacting with the instructor during a lecture, there are tens of students in the same classroom who either hesitate to engage in or outright dislike such interactions. While there are myriad explanations for this, ranging from social shyness and the fear of appearing clueless to one's peers to the inability to quickly articulate one's challenges with the presented material [13]–[16], the end result of the "fallacy of behavioral equivalence" is that the instructor can seldom, if ever, take real-time remedial actions to correct students' understanding of the course material.

The fallacy of learning equivalence

Much of the learning in engineering classes takes place through problem solving. In most—if not all—engineering classes, however, the lecture format leaves little time for in-class problem solving. Engineering instructors try to overcome this limitation of the lecture format by assigning homework and practice problems to students. In doing so, the instructors make an implicit assumption that all students are equally capable of learning through out-of-class problem solving.

But this too is a fallacious assumption. Consider, for example, what happens when a student gets stuck on an assigned problem due to conceptual challenges. The common thinking is that such students would reach out to the teaching staff (instructor, teaching assistants, etc.) for help. This, however, does not happen for a great majority of students due to reasons that range from their inability to approach the teaching staff during the assigned hours to the inefficacy of e-mail as a medium for discussing mathematical concepts [17], [18]. (Engineering instructors, for example, can often be heard complaining about students' lack of participation in office hours discussions.) The unfortunate consequence of this "fallacy of learning equivalence," especially in large classes, is that students' learning begins to go out of lockstep with each passing lecture.

Contemporary alternatives to the lecture format

The limitations of the lecture format, especially in the case of engineering education, are well known to the academic community. Several alternatives have been proposed and experimented with in recent years to overcome these limitations. Three modes of instruction that particularly stand out among these alternatives are 1) project-based learning, 2) a (massive open) online course, and 3) flipped classroom. While each one of these alternatives has its own sets of pros and cons, I decided to experiment with the flipped classroom based on the following observations.

- Project-based learning helps students gain a deeper understanding of the course material by presenting them with a real-world problem and guiding them toward a possible solution in a structured manner [19]–[22]. It is perhaps one of the most

FIGURE 1. One of the active learning classrooms at Rutgers University; facilities such as this are often recommended in education circles for use in project-based learning, flipped learning, etc. (Photo courtesy of Rutgers Digital Classroom Services.)

engaging modes of instruction and research has shown it to be highly effective in overcoming limitations of the lecture format [22]. Project-based learning, however, has its own set of challenges when it comes to its adoption for engineering education. It is not straightforward to design a project-based learning curriculum for the majority of core engineering courses. Further, project-based learning requires specialized active learning classrooms (see Figure 1), which are typically in short supply on most university campuses. Finally, the human resource overhead (in terms of man-hours and student–faculty ratio) associated with project-based learning deters cash-strapped academic departments with large student enrollment from fully embracing it as a scalable alternative to traditional lecturing.

- Online courses, in general, and massive open online courses (MOOCs), in particular, are often put forth as scalable alternatives to the lecture format [23], [24]. The single biggest advantage of online courses is that video archiving of instructor's presentations enables students to digest new material at their own pace by pausing, rewinding, and fast-forwarding parts of videos. Strictly speaking, however, online courses (MOOCs or otherwise) are pedagogically near-identical twins of lecture-based courses. Similar to the lecture format, they revolve around the passive transfer of knowledge from the instructor to students and implicitly assume the behavioral and learning equivalence of students. In fact, if anything, the lack of face-to-face interactions with the instructor only make it more challenging for some students to achieve the learning objectives of online courses. And the astronomical drop-out rates of MOOCs [25], [26] seem to confirm this impression that online courses are pedagogically challenging for all but the most resolute of students.
- Flipped classrooms (see "Anatomy of a Flipped Classroom"), popularized in K–12 education by the advent of Khan Academy [27], appear to strike somewhat of a balance between the high-overhead of project-based learning and the overly passive nature of online courses in engineering education. Similar to online courses, a flipped classroom makes use of video-based instructions that allow students the flexibility of revisiting key concepts at later stages in the course. Similar to project-based learning, a flipped classroom uses class time for activities that not only help students recognize deficiencies in their understanding of course material but also enable the instructor to take real-time remedial steps that can address these deficiencies. It is no surprise then that flipped classrooms have been adopted by a number of engineering instructors in recent years [28]–[33]. Notwithstanding these adoptions, the conventional wisdom among engineering instructors has been that a flipped classroom—similar to project-based learning—is not scalable to core engineering courses that enroll hundreds of students. There are two main reasons for this perception. First, it is a common belief that flipped offerings also require active learning classrooms. Second, positive learning outcomes in flipped classrooms are often linked to low student–faculty ratios. The fact that flipped classrooms in engineering education have mostly been adopted for small (sometimes elective) classes seems to strengthen this perception. Among the documented flipped classrooms in electrical engineering, [29], [30], and [31] had 30, 115, and 40 students, respectively.

Flipping digital signal processing at Rutgers University

Background and motivation

ECE 346: Digital Signal Processing is a required course at Rutgers for students majoring in electrical engineering. It is offered every year in the spring semester, with an average final enrollment of more than 100 students in the last five years. Traditionally, more than two-thirds of the students enrolling in this course are juniors who took ECE 345: Linear Systems and Signals in the immediately preceding semester, while the rest are seniors who did not or could not enroll earlier in the signal processing course for various personal or academic reasons. I have been teaching this course since spring 2012, with my first offering very much in the mold of traditional lecture and chalkboard format. This first offering would be considered a success by most

academic standards; the course quality received an average rating of 4.33 (out of five) from 56% of the enrolled students, and there were more than a handful of students who had truly mastered the course material by the end of the semester. Despite its seeming success, this first offering also laid bare to me many of the limitations of the lecture format, especially in relation to large core courses. In particular, the struggles of students who did not conform to the assumptions of the lecture format (see the section "The Case Against the Lecture Format") were all too palpable during the semester.

I made several tweaks to my first offering in the ensuing semesters in an attempt to make these offerings more equitable to students. These tweaks included experimenting with presentation slides in lieu of chalkboard text, video archiving of class lectures, grade-based incentives for class participation, and different attendance policies. Some of these tweaks appeared to be helpful to students' learning (e.g., video archiving), while other tweaks seemed to have either little effect (e.g., mandatory attendance) or negative effect (e.g., presentation slides). None of the tweaks seemed to

Anatomy of a Flipped Classroom

In a traditional classroom, the instructor transfers knowledge to the students by delivering weekly lectures during assigned class periods. The students are then expected to master the covered material outside the classroom by working on assigned homework exercises and reaching out to the instructor during assigned office hours for any clarifications. Traditional classrooms, unfortunately, do not work equally well for all students (see the section "The Case Against the Lecture Format"). A flipped classroom (also referred to as an *inverted classroom*) literally flips the traditional learning paradigm on its head (see Figure S1) [6]–[12]. Specifically, the knowledge transfer component of the course in a flipped classroom is moved outside the class; this typically involves the use of video lessons (see the section "Contemporary Alternatives to the Lecture Format"). The freed-up time during the assigned class periods is then used for carefully designed activities and collaborative exercises that help students master the course material. This "flipping" not only helps the students clarify any confusions in real time, but it also enables the instructor to personalize instructions to individual students based on their own gaps in understanding (see the section "Flipping Digital Signal Processing at Rutgers University").

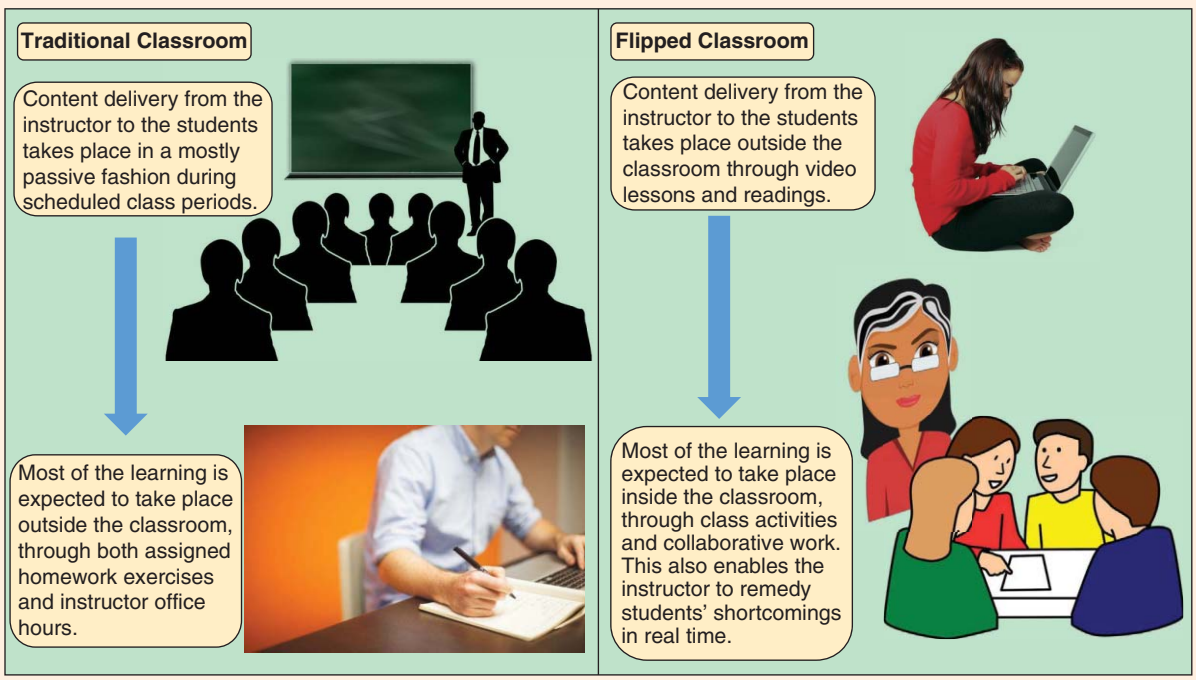


FIGURE S1. A side-by-side comparison of (a) a typical lecture-based classroom and (b) a flipped classroom. Because of the nature of these two modes of instruction, lecture-based learning and flipped learning are sometimes also referred to as *passive learning* and *active learning*, respectively.

directly confront the challenges of academic, behavioral, and learning variations among students. It was during this time, when I was exploring different means of teaching signal processing, that an interesting development took place. Prof. Van Veen taught flipped version of a senior-level elective signal processing class to 30 students at the University of Wisconsin-Madison in the fall of 2012 and shared his (highly positive) experience in [29]. The term *flipped clasroom* entered in my lexicon in 2013 as a result of [29], and I spent the next two years discussing with other educators (including the author of [29]) means by which large core engineering courses could possibly be flipped using minimal time, cost, and infrastructure overhead.

Ingredients of flipping on a shoestring

There were three major challenges that came to the fore when I carefully examined the possibility of flipping the mandatory junior-level signal processing class at Rutgers. First, and this is perhaps the most daunting aspect of flipping a course for any instructor, I needed a plan to create engaging video lessons in a cost- and time-effective manner. Second, and as noted by other instructors of flipped classes [29], [30], flipping a course for hundreds of students requires more than one person to guide students during in-class activities.

A general rule of thumb for the student–guide ratio in flipped classes is 20–30 students/guide, which means I needed a strategy to involve four to five additional guides in my flipped classroom without creating a budgeting crises for my department. The third challenge, often considered one of the biggest hurdles to the adoption of flipped learning for large engineering courses, is that the largest active learning classroom at Rutgers has a capacity of 90 students. Since enrollment in the Digital Signal Processing course at Rutgers often exceeded 100 students, I needed a plan that would enable students to reap the benefits of a flipped classroom in a lecture hall setting. The different ways in which I addressed these three challenges are described next.

Low-overhead video lessons

Short, self-contained video lessons are the key to creating a flipped classroom. But planning, recording, and producing professional-looking videos can overwhelm even the most committed of instructors. Being cognizant of the risks of overcommitting, I opted for an acceptable compromise between overhead and quality of the video lessons for my flipped offering. This compromise involved: 1) delivering lectures to students enrolled in my traditional offering of the signal processing class in spring 2015 using a pen tablet (Wacom Bamboo Tablet) connected to a Windows laptop and Microsoft OneNote, 2) capturing a laptop's screen using a screencasting software (Camtasia Studio 8) and recording voice using an external mic (Logitech HD Webcam), and 3) stitching, slicing, and deleting the recorded material using Camtasia Studio 8 to produce a set of 27 videos, each one of which covered a single topic and excluded classroom interactions and discussions with students. These videos, which are further divisible into subtopics of durations ranging from ten min to 30 min, are publicly available on my YouTube channel [34]. This piggybacking on traditional lecturing allowed me to limit the time overhead of these video lessons to an average of approximately 2.5 hours per video. (This figure excludes both the lecture preparation and the lecture delivery times since I would have spent this much time regardless as part of the spring 2015 offering.)

The monetary overhead of these video lessons was also quite manageable, enabling my department to absorb the entire cost; in particular, an equivalent system comprising a pen tablet, an external mic, and screencasting and video editing software can be built as of this writing for approximately $US400. This figure excludes the costs of a laptop and note-taking software, both of which are considered integral for today's educators.

Low-cost in-class assistants

While having a person assisting every 20–30 students for in-class activities is critical to the success of a flipped classroom, most universities cannot financially afford such a high ratio of students to teaching assistants. The junior-level signal processing class at Rutgers, for instance, has historically been assigned one graduate teaching assistant (GTA). To balance the needs for financial prudency and in-class assistants, I resorted to the use of peer learning assistants (LAs) for in-class activities. Specifically, I—along with the help of Rutgers Learning Centers—recruited five students from my previous (spring 2015) offering of the signal processing class to serve as LAs for in-class activities. Each one of these LAs spent two hours per week preparing for in-class activities and three hours per week assisting students during class times. These LAs were formally coached at the start of the semester in the art of pedagogy by Rutgers Learning Centers, and each one of them received a total of US$1,500 for the 14 weeks of instructions. Thus, for a meager monetary overhead of US$7,500 (split among the university and the department), my flipped offering resulted in a student–guide ratio of 22 (five LAs and one instructor for 133 enrolled students).

Flipping in a lecture hall

While an instructor should ideally have access to an active learning facility for a flipped offering [29], [30], the capital cost associated with construction of such facilities—especially the ones that can accommodate hundreds of students—means this is not always possible. I faced this very challenge for my flipped offerings at Rutgers. Rather than being deterred by this challenge, I retooled my flipped offerings for large lecture halls. This retooling involved 1) reserving a lecture hall for the flipped classroom whose capacity was at least twice the maximum expected course enrollment, 2) dividing the lecture hall into contiguous groups of three rows each, and 3) prohibiting students from sitting in the middle row of each group of rows. These empty middle rows

enabled the instructor and the LAs to freely roam around the lecture hall, be able to physically approach all students, and assist them during in-class activities (see Figure 2). While such a seating arrangement cannot be considered a replacement for an active learning facility, in which students themselves can also roam around and can utilize resources such as computers and writing boards (see Figure 1), mid- and end-of-semester feedback from students (see the section "Reflections on the Flipped Offering") suggested that the solution was an effective compromise between idealism and realism.

Course organization

My seminal flipped offering of the signal processing class physically met for 80 min each at 8:40 a.m. on Mondays and Thursdays. In addition, enrolled students were divided into three recitation groups, with each group attending one 80-minute recitation (led by the GTA) per week. There were three main categories of activities within this offering that fundamentally differentiated it from a traditional offering (see Table 1 for a bird's-eye view of these activities). These categories, referred to as home activities, in-class activities, and recitation activities in the offering's parlance, accounted for 29% of a student's final grade. To achieve the learning objectives of this offering, which included a comprehensive understanding of sampling theory, discrete-time processing of continuous-time signals, discrete Fourier transform, spectral analysis, and design of digital filters, I organized the three sets of course activities as follows.

Home activities

The category of home activities comprised tasks that students were required to complete outside the classroom. These tasks, the graded portion of which accounted for 7% of a student's final grade, were further subdivided into three groups. First, the students were regularly assigned video lessons, ranging in total duration from 30 to 70 min, and textbook reading that had to be watched and completed, respectively, before each class period. Second, each set of assigned video lessons and textbook reading was associated with an online assessment on a course management system (CMS) that the students had to complete by 7 a.m. on the day of the respective class. These online assessments comprised the simplest of short-answer, true-false, and multiple-choice questions and served as one of the main motivating factors for the students to watch the assigned videos and complete the assigned reading. There were two other aspects of the online assessments that gave students the opportunity to remedy some of the shortcomings in their understanding of the covered material. These involved giving ample time to the students to complete an online assessment (typically, an average of 3–5 min per question) and allowing students to retake an online assessment (with a different set of questions) in the case of unsatisfactory performance on the first attempt. My flipped offering in spring 2016 had a total of 20 online assessments, which accounted for 67% of the grade for home activities. The final group of tasks that constituted home activities mostly consisted of paper-and-pencil exercises meant to reinforce students' understanding of the course material; refer to Figure 3 for a graphical representation of home activities in my flipped classroom.

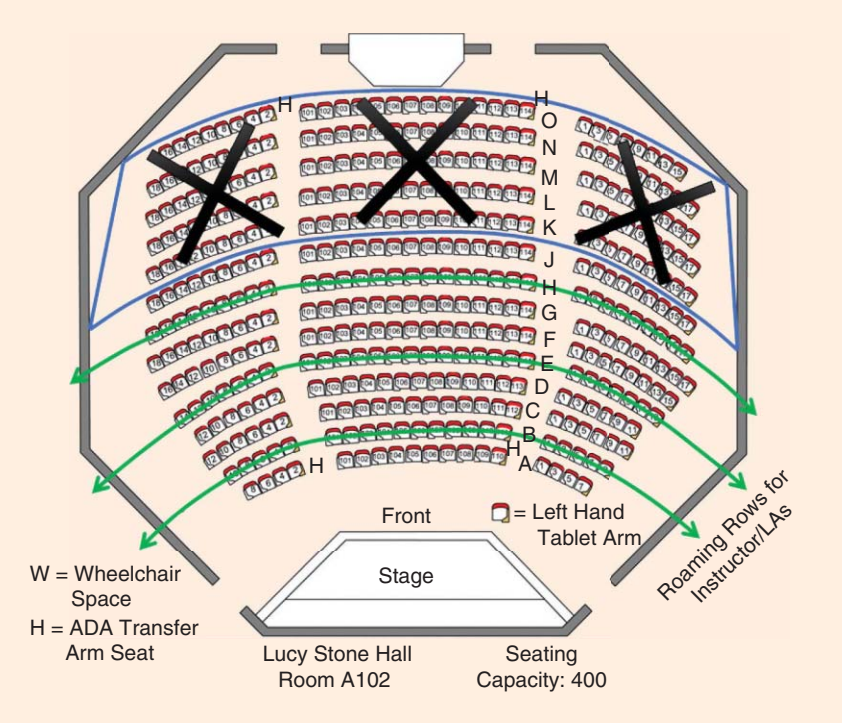


FIGURE 2. A possible seating arrangement in a lecture hall for students in a large flipped classroom. This arrangement, which I am using in my spring 2017 flipped offering at Rutgers, prohibits students from sitting at the very back of the hall (black crosses) and in three rows (green arrows), and enables instructor(s)/LA(s) to reach individual students by moving within the (green) restricted rows. This particular seating arrangement can accommodate up to 161 students, while it is scalable up to 257 students.

In-class activities

I divided each 80-min class period into two components. The first component, which typically lasted for 10–15 min, was used for a brief review of key concepts covered in the assigned video lessons. The second component, which covered the remaining class time and accounted for 15% of a student's final grade, comprised activities that helped students reflect on their understanding of the assigned video lessons and enabled me to take real-time remedial actions in response to widespread confusions. To this end, these activities

Table 1. A summary of the main activities that comprised my flipped offering in spring 2016.

Step Number	Activity Category	Activity Details	Grading Details
1–1	Home activity	Viewing of assigned YouTube video lessons (~30–70 min per class)	Ungraded
1–2	Home activity	Completion of assigned textbook reading (if applicable)	Ungraded
1–3	Home activity	Completion of online assessment (due by 7 a.m. on the day of each class)	~5% of the final grade
2–1	In-Class activity	Review of key concepts by the instructor (~10–15 min per class)	Ungraded
2–2	In-Class activity	Short polling questions (approximately two to five questions, with each worth two points)	15% of the final grade (25% points for an attempt)
2–3	In-Class activity	Paper-and-pencil problems (approximately one to three problems, with each worth four to 12 points)	
3–1	Home activity	Paper-and-pencil problems (approximately one to three problems assigned after some classes)	~2% of the final grade
4–1	Recitation activity	Problem solving by the GTA (~30–35 min and approximately three to five problems)	Ungraded
4–2	Recitation activity	Paper-and-pencil problems (approximately three to five problems, with each worth four to ten points)	7% of the final grade

were split into two categories: polling questions and paper-and-pencil exercises. The polling question part of in-class activities involved sequentially displaying short conceptual questions to students on a presentation slide and recording students' responses in real time using an online polling platform [see Figure 4(a)]. (I used the Poll Everywhere platform [35] in my class, which allows participants to respond using mobile devices.) The paper-and-pencil exercises part of in-class activities involved sequentially assigning longer problems [see Figure 4(b)] to students and collecting students' work on loose sheets of paper. A typical class period consisted of two to five polling questions and one to three paper-and-pencil exercises, with each polling question worth two points, each exercise worth anywhere between four and 12 points, and the students guaranteed 25% of the points for attempting an activity. I, along with the five LAs, helped the students during each ongoing activity by roaming around the lecture hall and providing

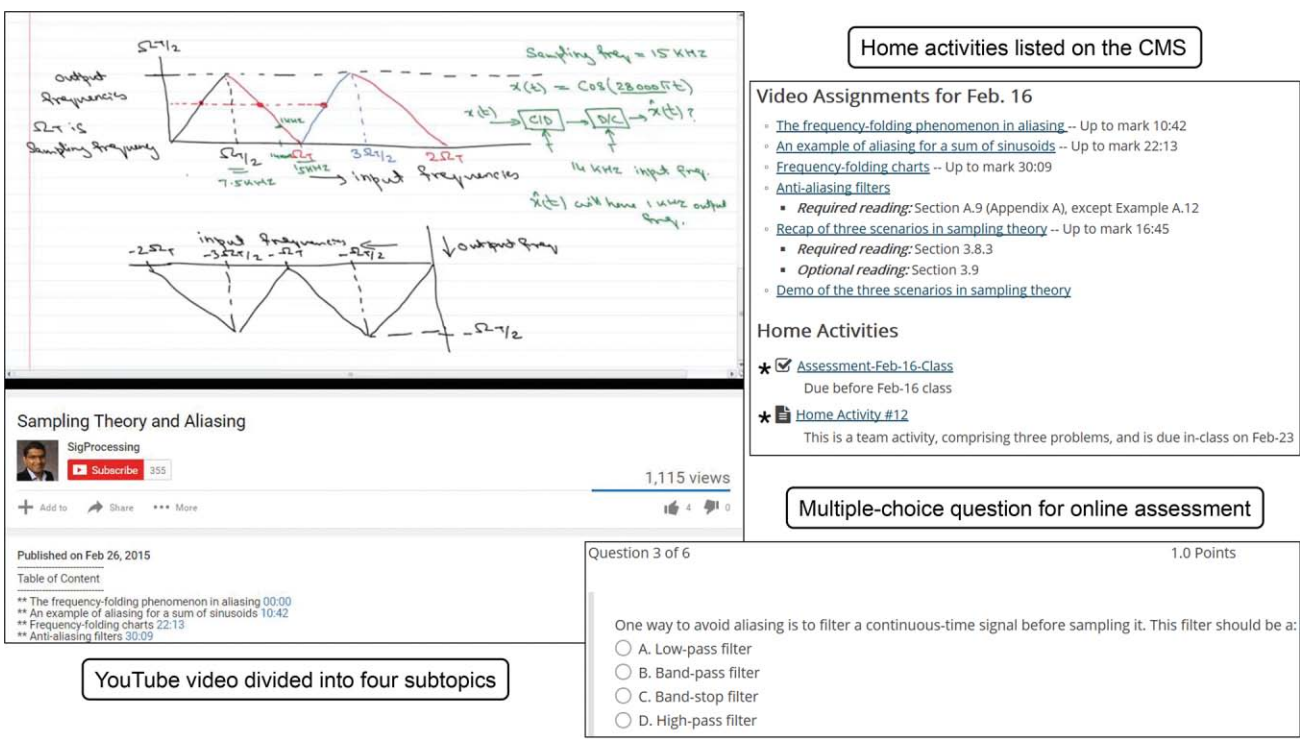


FIGURE 3. A graphical representation of the three main groups of tasks comprising home activities in my flipped classroom. The CMS screenshot corresponds to the spring 2017 flipped offering.

cues to struggling students. This looking over the shoulder of students and, in the case of polling questions, instantaneous access to students' responses [see, e.g., Figure 4(a)] gave me real-time insight into students' understanding of the covered material. This insight, which is one of the most important differences between a lecture-based offering and a flipped classroom, was then used to deliver a focused set of clarifying instructions to students at the end of each activity.

Recitation activities

Each one of the three recitation groups in the class attended one weekly 80-min recitation period led by the GTA. The activities in these weekly recitations were designed to enhance students' problem-solving skills. To this end, each recitation period was divided into two components. The first component, which typically lasted for 30–35 min, involved the GTA solving three to five problems on a chalkboard that reinforced the concepts covered in the last two sets of video lessons. The second component, which primarily distinguished the recitations in the flipped offering from those in a traditional offering, covered the remaining recitation period and accounted for 7% of a student's final grade. In this component, students were sequentially assigned three to five paper-and-pencil problems that specifically helped them master the mechanics of problem solving. This should be contrasted with the in-class paper-and-pencil exercises that focused on students' basic understanding of the course material. The students were given anywhere between four and ten min to solve each one of these problems on loose sheets of paper, with each problem worth anywhere between four and ten points. Further, the students were guaranteed 25% of the points for attempting a problem. The GTA, after assigning a problem to the students, roamed around the recitation room and helped students struggling with the problem. In addition, the students were encouraged to discuss the problems among themselves. Finally, the GTA capped off the assigned problems by collecting the students' work and having a brief discussion of solutions of the problems.

The rest of the flipped course's structure—apart from the aforementioned home, in-class, and recitation activities—followed a traditional offering, with the remaining 71% of a student's final grade divided among a prerequisite quiz, two in-class exams, a term project, and a final exam. There was, however, one additional aspect of my flipped offering that seemed to enhance students' learning experience. In the second week of the semester, after the enrollment transients died out, I divided my class into teams of three students (with at most two teams with four members each). The idea was that students on the same team would not only sit together during each class period and collaboratively work on in-class activities, but they would also work together on home activities comprising paper-and-pencil exercises. The success of this idea in terms of its impact on students' learning experience, however, depended on the creation of balanced teams. I accomplished this goal through the use of CATME Team-Maker tool [36], [37], which allows an instructor to gather various pieces of information from students and then assigns students to different teams according to the criteria and weighting specified by the instructor. (The CATME system has historically been free for use by the academic community. Starting 1 July 2017, however, there is expected to be a nominal license fee per unique student in an academic year that will help defray the system's annual maintenance costs.)

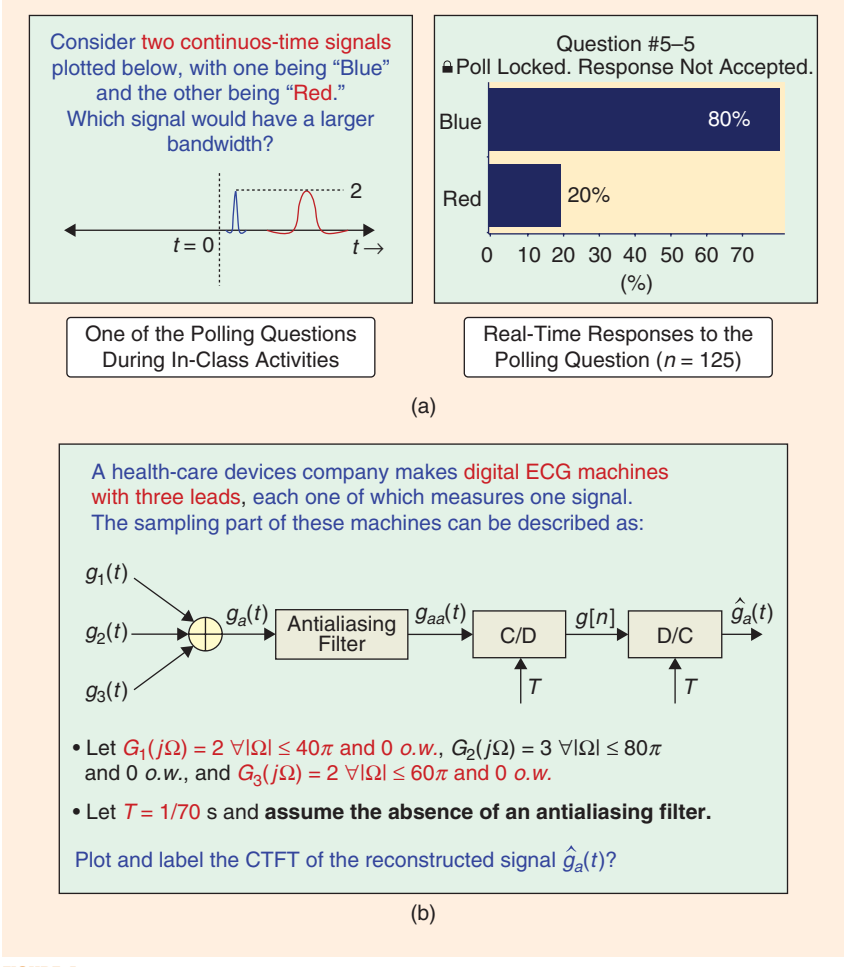


FIGURE 4. Illustrative examples of (a) a short conceptual question and students' responses to that question in my flipped classroom and (b) a longer paper-and-pencil exercise that the students solved within the class period.

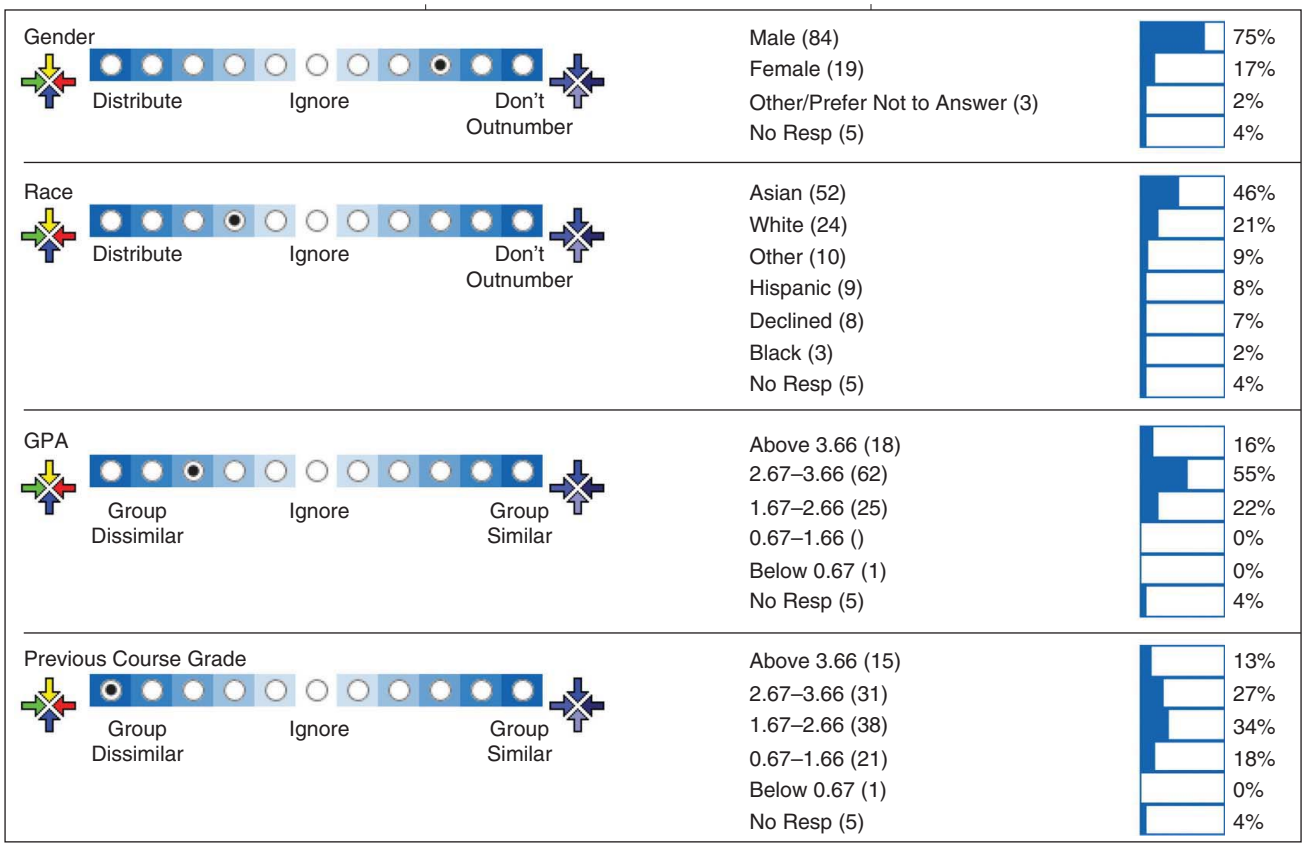


FIGURE 5. A partial screenshot of the CATME Team-Maker tool I used to distribute the students in my flipped classroom across different teams. This particular configuration of criteria and weighting corresponds to my spring 2017 offering with an enrollment of 111 students.

In particular, I configured the CATME Team-Maker tool such that the final set of teams brought together students with different levels of academic preparation, but similar (self-described) commitment levels, schedules, and class years; see Figure 5 for a sampling of the particular criteria and weighting I used for my spring 2017 flipped offering.

Reflections on the flipped offering

The discussion in the section "Flipping Digital Signal Processing at Rutgers University" makes it abundantly clear that my flipped offering was substantially different from a traditional lecture-based offering. But did this offering result in better learning outcomes for the students? Unfortunately, there are too many variables that affect students' learning abilities and a definitive answer cannot be given for this question without having the ability to control these variables. Some of these variables include students' academic preparation and command of prerequisite material, their learning styles, their work habits, and their intellectual abilities. Since I could not control any of these variables in my flipped offering, only anecdotal evidence from the perspectives of the instructor and the students can be provided to ascertain the effectiveness of the flipped offering.

Instructor's perspective

There are four data points from my perspective that seem to suggest that my seminal flipped offering was a success. First, the number of students attending each scheduled class period (see Figure 6) as well as the general body language of the students seemed to suggest the students were, on average, much more engaged in the flipped offering compared to my previous four traditional offerings. Second, the students' performance on in-class activities as well as the depth of their in-class queries suggested that the students internalized the course material better than in my previous offerings. Third, the sophistication of students' term projects in the flipped offering, on average, exceeded that of the projects in my traditional offerings. A possible explanation for this improvement is that students enrolled in the flipped classroom mastered the material better than in previous years. Finally, it used to be relatively straightforward for me in previous years to map students' numerical grades to letter grades. But the assignment of letter grades in the flipped offering became quite a chore for me due to the lack of significant gaps in the distribution of students' numerical grades. A possible explanation for this phenomenon, which has also been pointed out in [29], is that fewer students were being left behind in terms of their understanding as part of the flipped offering. In particular, the most noticeable aspects of my flipped offering—in comparison to the previous four

years—are the significantly higher percentage of "A" grades and the significantly lower percentage of "D" grades; see Figure 7 for the grade distributions of my 2012–2016 offerings.

Students' perspectives

Students' perspectives on this seminal flipped offering were formally obtained in two different manners. First, midway through the semester, I solicited anonymous feedback from the students on my flipped offering through a comprehensive survey of 24 questions. This survey, which had questions that ranged from numerical ratings (e.g., where a "1" meant "strongly disgree" and a "5" indicated "strongly agree" for "The traditional, lecture-based format of engineering education needs to be reformed") and multiple choice (e.g., "yes," "no," and "not sure" for "If given an opportunity, would you rather have ECE 346 in the traditional, lecture-based format?) to open ended (e.g., "What is one thing you would change if you were to offer ECE 346 as a flipped class?"), was completed by 63% of the enrolled students ($n = 84$). Among the survey takers, there were 75 students who had no prior experience with a flipped classroom. Figure 8 summarizes the responses of students to five main questions in the survey that reflected this cohort's opinion of engineering education, while Figure 9 summarizes students' responses to four key questions in the survey that can be interpreted as evaluation of my flipped offering. It can be seen from these two figures that, midway through the semester, an overwhelming majority of the students preferred the flipped classroom over the traditional lecture-based classroom. Students' responses to the open-ended questions in the survey shed some light onto a few of the reasons for this preference. According to one student, "I enjoy the overall aspect of watching the videos at home and then solidifying the information in class." Another student responded, "[It has given] me a chance to see what I do wrong when working out a problem […] during the class time[,] instead of working on homework and waiting a month to get it back and not knowing why I did what I did." And yet another student stated, "The flipped classroom method works better (in my opinion) because each student can go at his/her own pace."

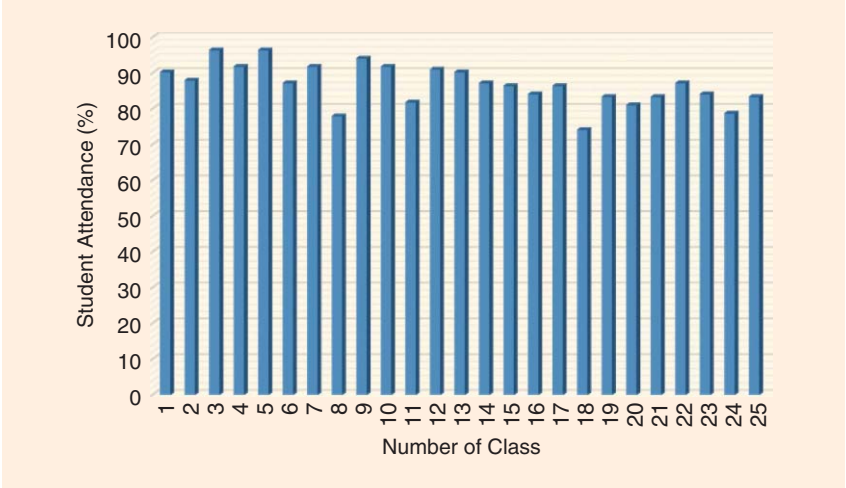


FIGURE 6. The percentage of students attending each class period of my spring 2016 offering (average attendance = 86%). These data correspond to the use of in-class activities grades as proxy for students' attendance, ignore the three class periods used for one review session and two in-class exams, and exclude the three students who withdrew from the course after the "drop" deadline.

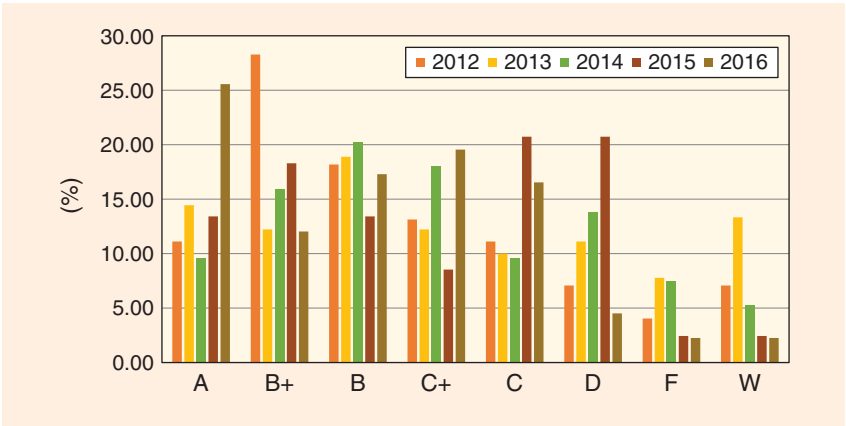


FIGURE 7. Grade distributions for my (spring) offerings of the signal processing class from 2012 to 2016. In terms of significant changes from year to year, presentation slides were used in lieu of chalkboard text for the 2013 offering, lecture archiving on YouTube was started from the 2014 offering, and a fully flipped class was offered in 2016.

While the open-ended responses of a number of students taking the midsemester survey validated my initiative, Figures 8 and 9 illustrate that not every student agreed with this initiative. There were, in particular, 13 students who would have preferred to enroll in a traditional lecture-based course (cf. Question 7 in Figure 9). The responses of these students to Questions 4, 5 and some of the open-ended survey questions helped explain their opposition to the flipped classroom. Nine of these 13 students responded with either "agree" or "strongly agree" to Question 4, while seven of them responded with either "agree" or "strongly agree" to Question 5. In terms of the open-ended questions, one of these students stated, "I feel that the flipped classroom is too much work for the amount of credits currently offered." Another student responded, "[…] the flipped classroom threw a curveball at me and I was slow to adapt. It certainly demands a higher time commitment […]" And yet another student stated,

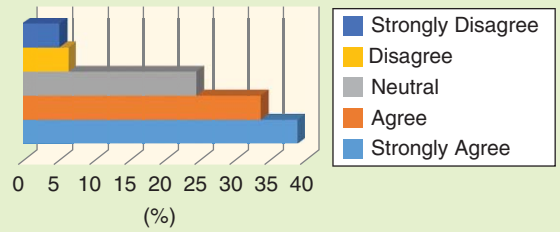
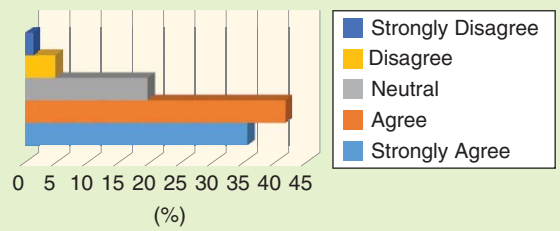
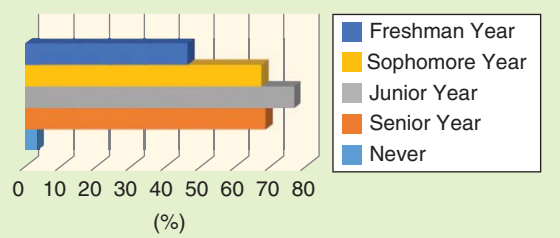
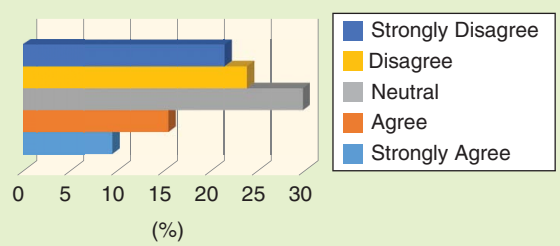
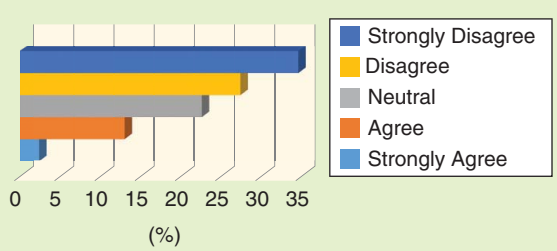
Number	Survey Question	Survey Response
1	The traditional, lecture-based format of engineering education needs to be reformed.	
2	A flipped classroom can be a good alternative to traditional, lecture-based format of engineering education.	
3	In engineering education, flipped classes should only be offered at the following levels (select one or more options).	
4	Flipped classes in engineering education should only be offered for small-sized (<20–25 students) courses.	
5	Flipped classes in engineering education should never be offered in lecture halls, even if it means reverting back to traditional, lecture-based instruction.	

FIGURE 8. Students' responses to five questions on a midsemester survey in spring 2016 that sought their personal opinions on the state of engineering education ($n = 84$).

"[I'm] doing very poorly in this course as of right now […] For this reason, I don't like it." These, and somewhat similar responses of a few other students, suggest that some of the students who preferred the lecture-based format might have done so for reasons other than pedagogical ones.

I obtained the next set of feedback on my flipped offering at the end of the semester as part of a Rutgers-administered anonymous course survey that helps students evaluate teaching effectiveness of the instructor and quality of the course. There were a total of 98 students (74% of the enrolled students) who responded to this survey. These students gave the flipped offering an average quality rating of 4.47, which is the highest quality rating I had received for my signal processing class. Note that I am all too familiar with the common refrain in some parts of the academy that the course quality (and instructor evaluation) ratings are inversely proportional to the amount of time students have to spend on the course. However, my flipped offering was nothing but a highly demanding class. One student, for example, noted in his end-of-semester survey, "The work load was extremely high[,] which helped with learning the material […]." Similarly, another student wrote, "The flipped classroom was an interesting experience, even though it was more work for students." In general, the feedback students provided through the end-of-semester survey corroborates findings of the midsemester survey and suggests better learning outcomes for a majority of the enrolled students. According to one student, "The abundance of examples and problems we did in class helped me understand the material more effectively than doing homework problems on my own. Despite being a class at [8:40] in the morning, I seldom felt tired or uninterested during the class." Similarly, another student responded, "The constant cycle of watching the videos, taking the quiz, reviewing in class, doing problems, and going to recitation to learn it again and do more problems was a fantastic process. It helped solidify every topic and drill it into my head […]." These are just few of the many survey responses that suggest students found the flipped offering to be both demanding and rewarding.

Concluding remarks

A flipped offering is a serious undertaking, both from the perspectives of the instructor and the students. In particular,

the amount of work and the additional resources required for successful offering of a flipped course can easily overwhelm the most dedicated of instructors. However, my experiences suggest that if one gradually transitions into a flipped offering and also adapts some aspects of a flipped classroom to the resource constraints of the offering university, then a flipped offering can be a truly rewarding experience for both the instructor and the students. But before that can be done, one has to be convinced of one thing: if some students are struggling academically in class, then it need not necessarily be due to their lack of effort. Once that realization sets in, only then can one go ahead and investigate pedagogical techniques that work better for those students.

When instructors try to answer the question of what works, they must be cognizant of the fact that the top-performing students cannot be used as a yardstick for success of a pedagogical style. Indeed, there are always going to be students in every class who would succeed regardless of the pedagogical techniques adopted in the class. But an instructor's duty is to reach out to all, and not just the top few, students. My seminal flipped offering has convinced me that an appropriately adapted flipped classroom is one way to reach out to more students. After having experimented with the flipped classroom and having seen the outcomes of this experiment, I am not planning to revert back to the traditional lecture-based format for my undergraduate signal processing class. Additionally, I hope that this article, along with the experiences of other engineering instructors, will be inspiring to people to begin their own quest for the elixir of scalable effective teaching.

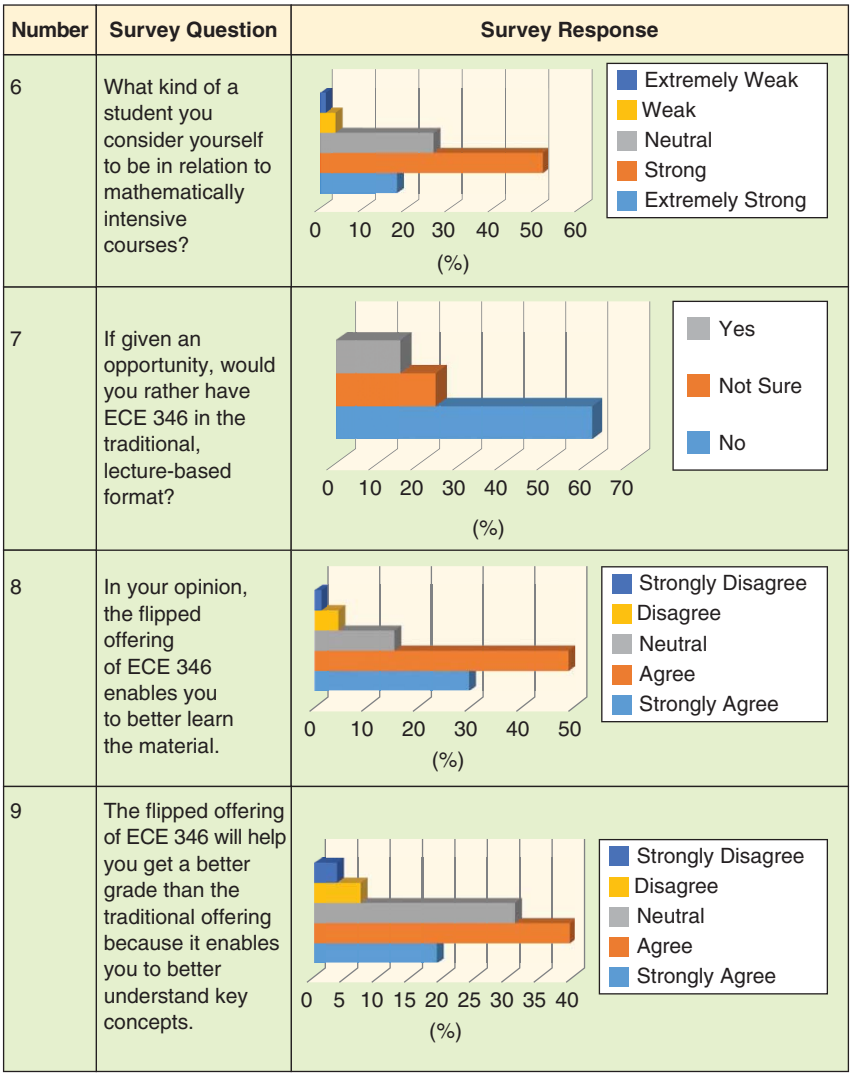


FIGURE 9. Students' responses to four questions on a midsemester survey in spring 2016 that can be interpreted as evaluation of my flipped offering ($n = 84$).

Acknowledgments

I owe the success of my seminal flipped offering both to discussions with numerous educators as well as support from several sources. In particular, I would like to acknowledge the support of the National Science Foundation CAREER Program, Rutgers Electrical and Computer Engineering, National Academy of Engineering Frontiers of Engineering Education, Rutgers Learning Centers, Rutgers Digital Classroom Services, and Rutgers Scheduling and Space Management. I would also like to thank Susan Albin, Helen Buettner, Mary Emenike, Lawrence Rabiner, Barry Van Veen, and Roy Yates for many helpful discussions, and the anonymous reviewers for several helpful comments.

About the author

Waheed U. Bajwa (waheed.bajwa@rutgers.edu) is with the Department of Electrical and Computer Engineering at Rutgers University in New Jersey. He received the 2015 National Science Foundation CAREER Award, whose education component supports his endeavors related to flipping learning.

References

[1] J. C. Scott, "The mission of the university: Medieval to postmodern transformations," *J. Higher Educ.*, vol. 77, no. 1, pp. 1–39, 2006.

[2] A. King, "From sage on the stage to guide on the side," *College Teaching*, vol. 41, no. 1, pp. 30–35, Jan. 1993.

[3] K. Powell, "Science education: Spare me the lecture," *Nature*, vol. 425, no. 6955, pp. 234–236, Sept. 2003.

[4] H. L. Lujan and S. E. DiCarlo, "Too much teaching, not enough learning: What is the solution?" *Adv. Phys. Educ.*, vol. 30, no. 1, pp. 17–22, Mar. 2006.

[5] H. G. Schmidt, S. L. Wagener, G. A. C. M. Smeets, L. M. Keemink, and H. T. van der Molen, "On the use and misuse of lectures in higher education," *Health Profession Educ.*, vol. 1, no. 1, pp. 12–18, Dec. 2015.

[6] J. Bishop and M. A. Verleger, "The flipped classroom: A survey of the research," in *Proc. ASEE Annu. Conf. Exposition*, Atlanta, GA, June 2013, pp. 23.1200.1–23.1200.18.

[7] G. Mason, T. R. Shuman, and K. E. Cook, "Inverting (flipping) classrooms—Advantages and challenges," in *Proc. ASEE Annu. Conf.*

Exposition, Atlanta, GA, June 2013, pp. 23.828.1–23.828.21.

[8] A Guide to the flipped classroom. The Chronicle of Higher Education (2015, Jan. 07) [Online]. Available: http://www.chronicle.com/article/A-Guide-to-the-Flipped/151039/

[9] J. L. Jensen, T. A. Kummer, and P. D. D. M. Godoy, "Improvements from a flipped classroom may simply be the fruits of active learning," *CBE Life Sci. Educ.*, vol. 14, no. 1, pp. 1–12, Mar. 2015.

[10] E. F. Gehringer, "Resources for "flipping" classes," in *Proc. ASEE Annu. Conf. Exposition*, Seattle, WA, June 2015, pp. 26.1336.1–26.1336.10.

[11] J. O'Flaherty and C. Phillips, "The use of flipped classrooms in higher education: A scoping review," *Internet Higher Educat.*, vol. 25, pp. 85–95, Oct. 2015.

[12] Flip Learning: Research, Reports, and Studies. Flipped Learning Network. [Online]. Available: http://flippedlearning.org/research-reports-studies/

[13] B. J. Limbach and W. L. Waugh, "Questioning the lecture format," *NEA Higher Educ. J. Thought Action*, vol. 20, no. 1, pp. 47–56, 2005.

[14] M. Freeman, P. Blayney, and P. Ginns, "Anonymity and in class learning: The case for electronic response systems," *Australasian J. Educ. Tech*, vol. 22, no. 4, pp. 568–580, 2006.

[15] J. R. Stowell and J. M. Nelson, "Benefits of electronic audience response systems on student participation, learning, and emotion," *Teaching Psychol.*, vol. 34, no. 4, pp. 253–258, 2007.

[16] C. R. Graham, T. R. Tripp, L. Seawright, and G. Joeckel, "Empowering or compelling reluctant participators using audience response systems," *Active Learn. Higher Educ.*, vol. 8, no. 3, pp. 233–258, 2007.

[17] Y. K. Kim and L. J. Sax, "Student–faculty interaction in research universities: Differences by student gender, race, social class, and first-generation status," *Res. Higher Educ.*, vol. 50, no. 5, pp. 437–459, 2009.

[18] W. Griffin, S. D. Cohen, R. Berndtson, K. M. Burson, K. M. Camper, Y. Chen, and M. A. Smith, "Starting the conversation: An exploratory study of factors that influence student office hour use," *College Teach.*, vol. 62, no. 3, pp. 94–99, 2014.

[19] P. C. Blumenfeld, E. Soloway, R. W. Marx, J. S. Krajcik, M. Guzdial, and A. Palincsar, "Motivating project-based learning: Sustaining the doing, supporting the learning," *Educ. Psychol.*, vol. 26, no. 3-4, pp. 369–398, June 1991.

[20] H. A. Hadim and S. K. Esche, "Enhancing the engineering curriculum through project-based learning," in *Proc. 32nd Annu. Frontiers in Education (FIE'02)*, Nov. 2002, vol. 2, pp. F3F.1–F3F.6.

[21] M. Frank, I. Lavy, and D. Elata, "Implementing the project-based learning approach in an academic engineering course," *Int. J. Tech. Design Educat.*, vol. 13, no. 3, pp. 273–288, Oct. 2003.

[22] J. S. Krajcik and P. C. Blumenfeld, "Project-based learning," in *The Cambridge Handbook of the Learning Sciences*, R. K. Sawyer, Ed. New York, NY: Cambridge Univ. Press, 2006, ch. 19, pp. 317–334.

[23] J. Bourne, D. Harris, and F. Mayadas, "Online engineering education: Learning anywhere, anytime," *J. Eng. Educ.*, vol. 94, no. 1, pp. 131–146, Jan. 2005.

[24] L. Pappano. (Nov. 2, 2012). The year of the MOOC. The New York Times. [Online]. Available: http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html

[25] D. F. O. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: Behavioural patterns," in *Proc. 6th Intl. Conf. Education and New Learning Technologies (EDULEARN'14)*, Barcelona, Spain, July 2014.

[26] T. A. Baran, R. G. Baraniuk, A. V. Oppenheim, P. Prandoni, and M. Vetterli, "MOOC adventures in signal processing: Bringing DSP to the era of massive open online courses," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 62–83, July 2016.

[27] Khan Academy. [Online]. Available: http://www.khanacademy.org/

[28] R. H. Rockland, L. Hirsch, L. Burr-Alexander, J. D. Carpinelli, and H. S. Kimmel, "Learning outside the classroom—Flipping an undergraduate circuits analysis course," in *Proc. ASEE Annu. Conf. Exposition*, Atlanta, GA, June 2013, pp. 23.854.1–23.854.8.

[29] B. Van Veen, "Flipping signal-processing instruction," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 145–150, Nov. 2013.

[30] M. L. Fowler, "Flipping signals and systems—Course structure & results," in *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing (ICASSP'14)*, Florence, Italy, May 2014, pp. 2219–2223.

[31] G. J. Kim, M. E. Law, and J. G. Harris, "Lessons learned from two years of flipping Circuits I," in *Proc. ASEE Annu. Conf. Exposition*, Seattle, WA, June 2015, pp. 26.1087.1–26.1087.12.

[32] M. G. Schrlau, R. J. Stevens, and S. Schley, "Flipping core courses in the undergraduate mechanical engineering curriculum: Heat transfer," *Adv. Eng. Educ.*, vol. 5, no. 3, Nov. 2016.

[33] J. R. Buck, K. E. Wage, and J. K. Nelson, "Designing active learning environments," *Acoustics Today*, vol. 12, no. 2, pp. 12–20, 2016.

[34] W. U. Bajwa. SigProcessing YouTube channel. [Online]. Available: http://www.youtube.com/user/SigProcessing

[35] Poll Everywhere. [Online]. Available: http://www.polleverywhere.com/

[36] R. A. Layton, M. L. Loughry, M. W. Ohland, and G. D. Ricco, "Design and validation of a web-based system for assigning members to teams using instructor-specified criteria," *Adv. Eng. Educat.*, vol. 2, no. 1, pp. 1–28, 2010.

[37] CATME System. [Online]. Available: http://info.catme.org/

SP

SP COMPETITIONS *(continued from page 150)*

Matthew E.P. Davies (matthew.davies@inesctec.pt) is a senior researcher at INESC TEC, Portugal. He is an IEEE Member and an associate editor for *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. He is responsible for the technical components of the competition and its materials and ran the competition alongside Craig T. Jin.

Patrizio Campisi (patrizio.campisi@uniroma3.it) is a professor at Roma Tre University, Rome, Italy. He chairs the Student Service Committee of the IEEE Signal Processing Society and served as the general chair of the 2015 Information Forensics and Security Technical Committee. He is the organizer for the SP Cup since the 2015 edition and the initiator of the Video and Image Processing Cup.

References

[1] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," in *Proc. 2nd ACM Int. Conf. Multimedia*, 1994, pp. 365–372.

[2] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, 1998.

[3] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Comput. Music J.*, vol. 29, no. 1, pp. 34–54, 2005.

[4] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordà, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. (2013). Roadmap for Music Information ReSearch. London: MIRES Consortium. [Online]. Available: http://mires.eecs.qmul.ac.uk/about.html

[5] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Queen Mary Univ., Centre for Digital Music, London, Tech. Rep. C4DM-TR-09-06, 2009.

[6] J. Oliveira, M. E. P. Davies, F. Gouyon, and L. P. Reis, "Beat tracking for multiple applications: A multi-agent system architecture with state recovery," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 10, pp. 2696–2706, 2012.

[7] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. "Madmom: A new Python audio and music signal processing library," in *Proc. 2016 ACM Multimedia Conf.*, 2016, pp. 1174–1178.

SP

LECTURE NOTES

Heinrich Edgar Arnold Laue

Demystifying Compressive Sensing

The conventional Nyquist–Shannon sampling theorem has been fundamental to the acquisition of signals for decades, relating a uniform sampling rate to the bandwidth of a signal. However, many signals can be compressed after sampling, implying a high level of redundancy. The theory of compressive sensing/sampling (CS) presents a sampling framework based on the rate of information of a signal and not the bandwidth, thereby minimizing redundancy during sampling. This means that a signal can be recovered from far fewer samples than conventionally required.

Relevance

When first exposed to a field, and especially one which challenges existing paradigms, it is often useful to start by gaining a high-level understanding of the principles underlying that field. This article employs a new set of analogies, illustrations, and numerical examples to provide intuitive explanations for the fundamental principles in CS. Armed with a feel for why CS makes sense, the interested reader may then proceed to more technical introductions to the field, such as [1]–[3].

Prerequisites

The body of this article requires knowledge of linear algebra, conventional sampling theory, basic probability, and basic optimization.

Digital Object Identifier 10.1109/MSP.2017.2693649
Date of publication: 11 July 2017

Problem statement

Sparsifying bases

Consider a discrete signal vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ of N samples taken at the conventional Nyquist rate of twice the signal bandwidth. A signal vector in a conventional sampling domain can be expressed in a different domain/basis for analysis or processing. For example, time-domain signals may be transformed to the Fourier domain to analyze their frequency content.

Consider a set of orthonormal basis vectors placed as columns in the transform matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1\, \boldsymbol{\psi}_2 \cdots \boldsymbol{\psi}_N]$. The signal vector $\mathbf{x}$ can then be expressed as a weighted sum of basis vectors [1]

$$\mathbf{x} = \sum_{n=1}^{N} s_n \boldsymbol{\psi}_n = \boldsymbol{\Psi}\mathbf{s}, \qquad (1)$$

where the $N \times 1$ vector $\mathbf{s}$ contains the coefficients of the signal in its new basis, found as the projection of $\mathbf{x}$ onto each of the basis vectors by the dot product $s_n = \langle \boldsymbol{\psi}_n, \mathbf{x} \rangle$, or $\mathbf{s} = \boldsymbol{\Psi}^T\mathbf{x}$. Each coefficient is represented by its own basis vector, which is separable from all others.

Sampling at the Nyquist rate guarantees perfect recovery of the original signal, suggesting that no fewer than N coefficients are required to fully describe the signal. However, a signal vector can often be expressed in a different basis where many coefficients are zero (or close to zero) [2]. The remaining K nonzero, or significant, coefficients are sufficient to fully describe the signal. When $\mathbf{x}$ is expressed in a sparsifying basis, it results in a K-sparse vector $\mathbf{s}$ with only $K \ll N$ significant coefficients. Sparsity in $\mathbf{s}$ implies redundancy in $\mathbf{x}$, since N samples represent a signal with effectively only K degrees of freedom. Here, the significant coefficients in the sparsifying basis can be seen as the concepts or information being conveyed by $\mathbf{x}$.

Figure 1 illustrates a two-dimensional (2-D) data point at (2, 1) in the standard basis. However, in the sparsifying basis shown, the data point is (2.2, 0), which has only one significant coefficient.

Many signals have sparsifying bases, a well-known fact in conventional compression, where a signal may be expressed in a sparsifying basis so only the largest coefficients can be retained [1].

Images are typically sparse in the discrete cosine transform (DCT) or wavelet bases [4], audio signals in the modified discrete cosine transform basis [5], magnetic resonance images in the spatial,

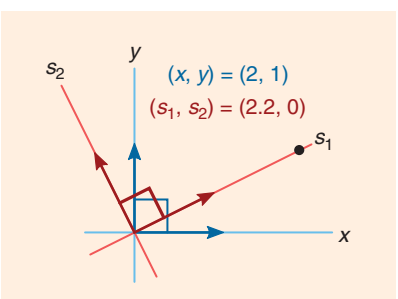


FIGURE 1. An illustration of a 2-D data point expressed in standard and sparsifying bases.

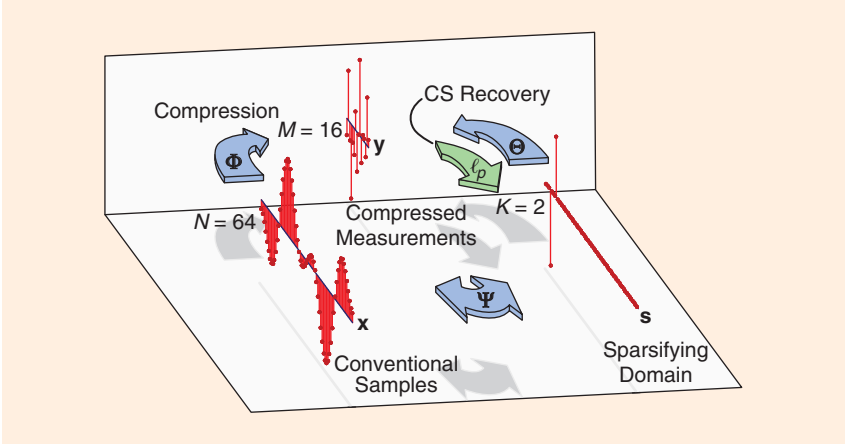


FIGURE 2. An example of a CS system with $N = 64$, $M = 16$, $K = 2$, random Gaussian $\mathbf{\Phi}$, and $\mathbf{\Psi}$ the inverse DCT matrix. The blue arrows represent matrix multiplications. Amplitudes are not to scale.

spatial finite differences, or wavelet domains [6], and sensor array data in direction of arrival [7].

Incoherence and compressive sampling

CS exploits redundancy to reduce the number of samples that must be taken to fully describe a signal. We have seen that redundancy can be quantified in terms of the number of significant coefficients in a sparsifying basis. CS aims to reduce the required number of samples without any prior knowledge of the signal, only the assumption that some sparsifying basis exists.

Imagine, as a first approach, that you decide to simply neglect some of the conventional Nyquist samples at random. The problem here is that some samples that you neglect may contain crucial information. For example, what if the signal contains spikes or sharp discontinuities? Neglecting samples in these areas would lead to significant loss of information.

For this approach to work, it would be necessary for the information to be distributed evenly over all of the conventional samples, so that no one sample conveys significantly more information than the rest. While the signal should be sparse in some other basis, it should certainly not be sparse in the domain in which it is sampled.

The incoherence between two domains expresses the idea that a vector which is sparse in one domain will be nonsparse in the other and occurs when the basis vectors between the domains are dissimilar [2]. Consider the time and frequency domains, where the time domain is represented by the standard basis and the frequency domain by the Fourier basis. A single-frequency component will result in a sine wave with most time samples being significant. Similarly, a time-domain impulse can only be represented by a multitude of frequency components.

The bottom of Figure 2 shows an example where $\mathbf{\Psi}$ is the inverse DCT matrix. The blue arrow represents a matrix multiplication with $\mathbf{\Psi}$ or its inverse, depending on the direction. With only two significant coefficients in $\mathbf{s}$, almost all conventional samples in $\mathbf{x}$ are significant.

If the bases of $\mathbf{x}$ and $\mathbf{s}$ are incoherent, and if $\mathbf{s}$ is sparse, then the information of interest (the K significant coefficients in $\mathbf{s}$) will be distributed over all N samples in $\mathbf{x}$. Since this information is comparatively little ($K \ll N$), neglecting some of the samples in $\mathbf{x}$ is unlikely to lead to a significant loss of information. (See "Analogy 1—Listening With Half an Ear.") Also, the choice of which samples to neglect becomes almost arbitrary, as long as enough are kept [2]. Aliasing is not a problem when the samples that are kept are not spaced uniformly; therefore, samples are typically neglected at random [2].

So far we require sparsity in $\mathbf{s}$, and nonsparsity in $\mathbf{x}$, which is met when there is incoherence between the conventional and sparsifying bases. But what happens if $\mathbf{x}$ itself is sparse? Redundancy is still present, but neglecting samples at random will result in a loss of information. The solution is to transform the signal $\mathbf{x}$ to an intermediate domain before sampling—a domain in which the signal is nonsparse, and that is incoherent with the sparsifying domain.

How will we select such an intermediate domain? An interesting fact is that a random basis is highly likely to be incoherent with almost any other basis [2]. This means that without prior knowledge of the sparsifying domain, we may transform the signal to a random domain and be sure that the incoherence requirement will be met. In this way we can design universal sampling schemes requiring only the assumption that some sparsifying basis exists. No knowledge of the sparsifying basis is required during sampling, only during signal reconstruction.

How is CS implemented? The simultaneous process of transforming to an intermediate domain and neglecting transformed samples may be described by the system $\mathbf{y} = \mathbf{\Phi x}$, where $\mathbf{y}$ is the $M \times 1$ vector of compressed samples (or measurements), and $\mathbf{\Phi}$ is the rectangular $M \times N$ sensing matrix, with $K < M < N$. The rows $\{\boldsymbol{\phi}_m\}_{m=1}^{M}$ describe the basis vectors of the intermediate domain, where only M out of N basis vectors are used. Each measurement is found as the projection of $\mathbf{x}$ onto the corresponding basis vector as $y_m = \boldsymbol{\phi}_m \mathbf{x}$. Figure 2 illustrates how a random $\mathbf{\Phi}$ is used to compress $\mathbf{x}$ into the measurement vector $\mathbf{y}$.

Each measurement y_m is a unique weighted combination of all of the elements in $\mathbf{x}$, or $y_m = \sum_{n=1}^{N} \phi_{m,n} x_n$. Notice how this ensures that the information in $\mathbf{x}$ is distributed over all of the measurements in $\mathbf{y}$. The processes of weighting and summation are used for hardware implementations of the sensing matrix. This is illustrated in Figure 3(a) for $N = 4$ and $M = 2$.

While a form of CS may be implemented in software to compress Nyquist-rate signals, the real power of CS lies in developing new hardware-based sub-Nyquist sampling schemes.

Figure 3(b) illustrates a CS application, the single-pixel camera, where an

Analogy 1—Listening With Half an Ear

Three journalists are taking notes at a press conference. The first is inexperienced and quite naïve; so afraid of missing something important, he frantically writes down every word being said. The second journalist is more experienced and while also listening attentively to every word, he interprets what is said and summarizes the facts concisely in his notes. The third journalist is quite lazy; not listening attentively at all, he only picks up every second or third word being said. The experienced journalist sees the lazy one daydreaming and is greatly surprised afterward to find their notes almost identical. "How did you get all of the facts, when you were clearly only listening with half an ear?" he asks. "Did you not know," the lazy one replies, "that the speaker is known to waffle, using ten words to convey a single concept? I am not likely to miss anything important when I know he could not be saying very much, though his words may be many."

The naïve journalist represents conventional Nyquist-rate sampling, with words analogous to samples. This scheme makes no effort to interpret what is being sampled; the sampling rate is based purely on the signal bandwidth. The experienced journalist represents conventional signal compression. Sampling is still done at the Nyquist rate, but the system interprets the sampled signal and expresses it more concisely for storage or transmission. The lazy, or rather, the efficient journalist represents CS. This scheme samples well below the Nyquist rate by assuming that the unique concepts or information being conveyed is little, and that this information is distributed over the many conventional samples so that missing a particular sample is unlikely to lead to significant loss of information.

N-pixel image is reconstructed from $M < N$ measurements [1], [4]. An image is projected onto a digital micromirror device (DMD), which is an array of tiny mirrors, each representing a pixel. For each measurement y_m, the micromirrors are randomly set to either reflect light toward a collecting lens or away from it. This is the process of weighting (by zero or one), with the weights for each measurement taken from the corresponding row in $\mathbf{\Phi}$. The lens then collects the rays from the DMD and concentrates it onto a single photodiode—the process of summation. After the first measurement, new weights are programmed into the DMD for the next measurement.

Solution

Equivalent systems and design requirements

So far we have seen that CS is possible when a signal has a sparse representation in some domain which is incoherent with the sampling domain. To understand the problem in more detail, it is useful to consider it from a slightly different angle. Consider the effect of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ not separately, but together in the following equivalent formulation of the CS problem:

$$\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Phi \Psi s} = \mathbf{\Theta s} \tag{2}$$

where $\mathbf{\Theta}$ is the $M \times N$ compressed transform matrix. The columns $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$ are vectors in an overcomplete basis, with more vectors than dimensions. Each vector still represents a coefficient in $\mathbf{s}$, but the set cannot be orthogonal and there will be some similarity between the vectors.

The goal is to recover $\mathbf{s}$ from $\mathbf{y} = \mathbf{\Theta s}$. Notice that the system is underdetermined since there are fewer equations than unknowns. It would generally be impossible to determine $\mathbf{s}$ uniquely since we do not have sufficient information about it from $\mathbf{y}$. However, the assumption of sparsity in $\mathbf{s}$ is the additional information we require to obtain a unique solution. Out of the infinite number of possible solutions, we choose to consider only the sparsest solution, i.e., the

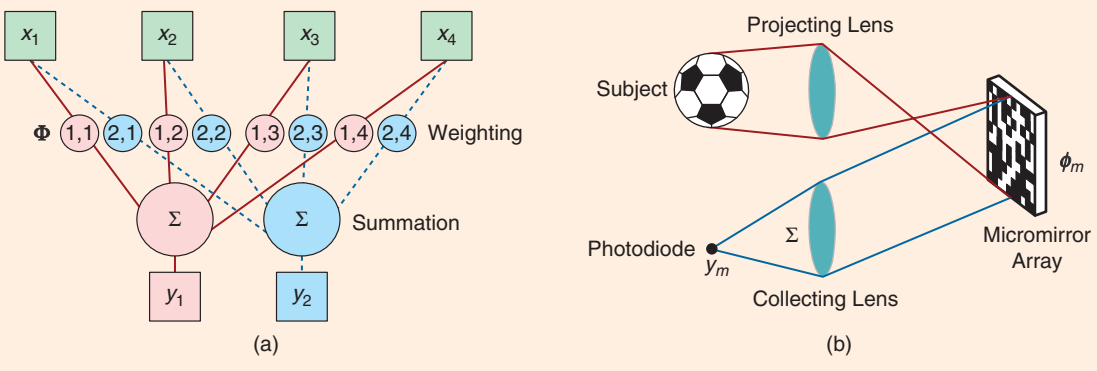


FIGURE 3. (a) A general illustration of a sensing matrix implementation ($N = 4, M = 2$). (b) A single-pixel CS camera (adapted from [1]).

Analogy 2—Filling in the Gaps

Imagine playing a game where a friend chooses a word and reveals the letters and their positions one by one, at random, until you correctly identify all of the letters. You have gotten as far as "co–pr–ss." You see that the answer must be "compress," and that the missing letters are "m" and "e." How did you identify these two letters? You did not see the letters as unrelated parts to be identified on their own but saw them as collectively conveying a single concept—a word.

CS is analogous to such a game, where letters represent samples. Instead of seeing samples as unrelated to each other, CS identifies the concepts the samples are collectively conveying. These concepts are few compared to the number of samples, like many letters convey a single word. Some samples required by the Nyquist theorem are missing, but by identifying the concepts being conveyed, the CS algorithm is able to fill in the missing samples.

On the other hand, imagine playing the game without the assumption that the letters form a word. This significantly increases the complexity of the problem, since any combination of letters is a possible solution. Without any vocabulary to draw from, there is no way to identify the missing letters and you will have to wait until all of the letters have been revealed.

Conventional sampling is like the latter scenario. It does not interpret what the samples are conveying but treats each sample as an individual concept to be identified correctly. All of the Nyquist samples must be taken; there is no way of filling in any gaps.

sparsest vector $\mathbf{s}$ that satisfies $\mathbf{y} = \mathbf{\Theta s}$. We are applying a form of Occam's razor: out of all of the possible explanations for the measured data, we assume that the simplest (sparsest) one must be correct. (See "Analogy 2—Filling in the Gaps.")

Figure 2 shows the relationship among $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Theta}$. The arrow labeled *CS recovery* indicates that, using a CS algorithm, we are able to go upstream by finding the length-N vector $\mathbf{s}$ from the smaller length-M vector $\mathbf{y}$. Having $\mathbf{s}$, we can find the original vector $\mathbf{x}$ through $\mathbf{\Psi}$.

Consider the following example. An underdetermined system in the form $\mathbf{y} = \mathbf{\Theta s}$ is given by (3), shown in the box below, where a random Gaussian sensing matrix $\mathbf{\Phi}$ with zero mean and variance $1/N$ and the inverse DCT transform matrix $\mathbf{\Psi}$ have been used to obtain $\mathbf{\Theta} = \mathbf{\Phi\Psi}$.

Assuming for now that we know which coefficients in $\mathbf{s}$ are significant, notice how the system may be reduced to the equivalent overdetermined subsystem

$$\begin{bmatrix} -0.357 \\ 0.612 \\ 0.137 \end{bmatrix} = \begin{bmatrix} -0.293 & -0.127 \\ 0.088 & 1.048 \\ -0.069 & 0.412 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} \tag{4}$$

in the form $\mathbf{y} = \mathbf{\Theta}'\mathbf{s}'$, where the columns in $\mathbf{\Theta}$ corresponding to the nonzero elements in $\mathbf{s}$ have been extracted to give the submatrix $\mathbf{\Theta}'$, and $\mathbf{s}'$ contains the nonzero coefficients in $\mathbf{s}$.

Under which conditions can $\mathbf{s}$ be recovered from $\mathbf{y}$? Notice first that the columns in $\mathbf{\Theta}'$ should be linearly independent, since the coefficients in $\mathbf{s}'$ will not be separable if their vectors can be expressed as a linear combination of the vectors corresponding to the other coefficients. This condition would have sufficed if we knew the K locations beforehand. However, we do not, and must, in addition, ensure that there is only one K-sparse solution to choose—a unique solution.

Suppose for the sake of contradiction that two distinct K-sparse solutions $\mathbf{s}$ and $\hat{\mathbf{s}}$ exist, such that $\mathbf{y} = \mathbf{\Theta s} = \hat{\mathbf{y}} = \mathbf{\Theta}\hat{\mathbf{s}}$, or $\mathbf{\Theta}(\mathbf{s} - \hat{\mathbf{s}}) = \mathbf{0}$. Then the difference vector $\boldsymbol{\delta} = \mathbf{s} - \hat{\mathbf{s}}$ is at most $2K$-sparse [8]. By definition, the equivalent subsystem $\mathbf{\Theta}'\boldsymbol{\delta}' = \mathbf{0}$ has a nontrivial solution if and only if the $2K$ or less columns in $\mathbf{\Theta}'$ are linearly dependent. Conversely, if these columns are linearly independent, $\mathbf{\Theta}(\mathbf{s} - \hat{\mathbf{s}}) = \mathbf{0}$ cannot be satisfied and no more than one K-sparse solution can exist. To guarantee a unique solution for any combination of K or less significant coefficients in $\mathbf{s}$, all subsets of $2K$ columns in $\mathbf{\Theta}$ must be linearly independent [8].

The restricted isometry property (RIP) goes a step further and considers whether submatrices with up to $2K$ columns are nearly linearly dependent, by placing bounds on the conditioning of these submatrices [9]. This ensures robustness in the presence of noise, since small perturbations may produce large errors when solving a nearly linearly dependent system.

Compressed transform matrices $\mathbf{\Theta}$ with independent and identically distributed (i.i.d.) random entries have been shown to meet the RIP criterion with high probability. Examples include Gaussian matrices with zero mean and variance $1/M$, and Bernoulli matrices with equiprobable $\pm 1/\sqrt{M}$ entries [2]. For practical implementations, $\mathbf{\Phi}$ may be chosen in the same way as $\mathbf{\Theta}$ (replacing M with N), and the resulting $\mathbf{\Theta}$ will still be able to meet the criteria for arbitrary choice of $\mathbf{\Psi}$ [2]. As a result, the incoherence requirement between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ will also be satisfied.

$$\begin{bmatrix} -0.357 \\ 0.612 \\ 0.137 \end{bmatrix} = \begin{bmatrix} -1.036 & -0.293 & -0.127 & -0.503 & -0.127 \\ -0.844 & 0.088 & -0.385 & 0.105 & 1.048 \\ 0.241 & -0.069 & -0.444 & 0.733 & 0.412 \end{bmatrix} \begin{bmatrix} 0 \\ 1.0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix} \tag{3}$$

Minimizing sparsity—perfect recovery algorithms

We have seen how it is possible for a CS system to preserve the information in $\mathbf{x}$ from only the measurements $\mathbf{y}$. But how will one go about recovering $\mathbf{x}$ from $\mathbf{y}$?

In developing a CS recovery algorithm, our aim is to obtain a unique solution to an underdetermined system that is the sparsest of all solutions. This can be formulated as

$$\min \| \tilde{\mathbf{s}} \|_0 \quad \text{subject to} \quad \mathbf{y} = \boldsymbol{\Theta}\tilde{\mathbf{s}}, \quad (5)$$

where $\| \cdot \|_0$ is the ℓ_0-norm defined as the number of nonzero elements in a vector. The problem reads: "Minimize the number of nonzero elements in $\tilde{\mathbf{s}}$, subject to $\tilde{\mathbf{s}}$ being a possible solution to the system." The general ℓ_p-norm is defined as $\| \mathbf{s} \|_p = \sqrt[p]{\sum_{n=1}^{N} |s_n|^p}$.

To illustrate, consider the following combinatorial ℓ_0-minimization algorithm:

for $k \leftarrow 1$ **to** $M - 1$

for all combinations of k out of N coefficient locations in $\tilde{\mathbf{s}}$

Find the least-squares solution $\hat{\mathbf{s}}'$ to $\mathbf{y} = \boldsymbol{\Theta}'\tilde{\mathbf{s}}'$

if $\hat{\mathbf{y}} = \boldsymbol{\Theta}'\hat{\mathbf{s}}' = \mathbf{y} \rightarrow$ **break.**

This algorithm tries all possible combinations of sparse coefficient locations, starting with a single one, until it finds an exact least-squares solution to the subsystem.

Consider the system in (3) with $\mathbf{s}$ unknown. The first try would assume that only s_1 is significant and solve the resulting subsystem, giving the least-squares solution $s_1 = -0.062$. This gives $\hat{\mathbf{y}} = [0.064, 0.052, -0.015]^T \neq \mathbf{y}$, which is not an exact solution.

Eventually we reach the combination (s_2, s_5), which leads to the subsystem in (4) with $\mathbf{s}'$ unknown, for which the least-squares solution is $(s_2, s_5) = (1.0, 0.5)$, giving $\hat{\mathbf{y}} = [-0.357, 0.612, 0.137]^T = \mathbf{y}$, an exact solution. This also happens to be the first combination that gives an exact solution, as expected. While there are many more solutions, the algorithm accepts this one as correct since it is sparsest. If desired, we can now calculate $\hat{\mathbf{x}} = \boldsymbol{\Psi}\hat{\mathbf{s}}$.

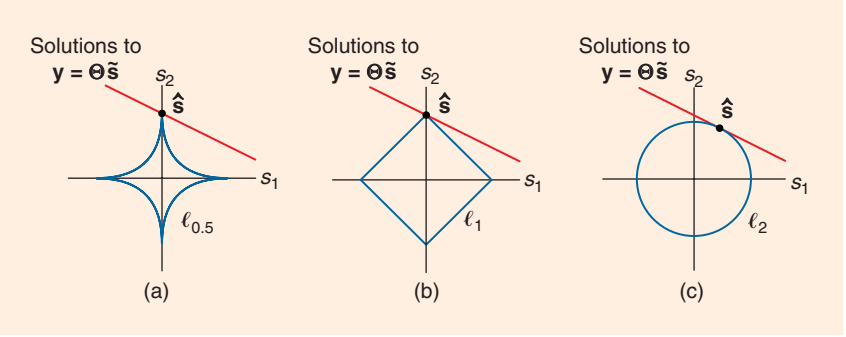


FIGURE 4. An illustration of ℓ_p-norms for (a) $p = 0.5$, (b) $p = 1$, and (c) $p = 2$ (Euclidean norm). Blue lines occur where the vector $\tilde{\mathbf{s}}$ has constant ℓ_p-norm. Any point on a red line is a possible solution.

Unfortunately, a combinatorial ℓ_0 algorithm is computationally infeasible for problems of practical sizes [1]. For a practically feasible recovery algorithm we use the ℓ_p-norm formulation

$$\min \| \tilde{\mathbf{s}} \|_p \quad \text{subject to} \quad \mathbf{y} = \boldsymbol{\Theta}\tilde{\mathbf{s}}, 0 < p \leq 1, \quad (6)$$

which can be solved using a variety of efficient optimization algorithms [10]. The ℓ_1 problem can be recast as a linear program [3], [8], and requires around $O(K \log(N/K))$ measurements when using a random sensing matrix [2]. For $0 < p < 1$ the problem is nonconvex with multiple minima, but local optimizers nevertheless perform well, particularly when $p = 0.5$ [10].

To visualize why the ℓ_p-norm favors sparsity when $0 < p \leq 1$, consider the graphs in Figure 4 for $p = 0.5$, $p = 1$, and $p = 2$. The ℓ_p balls shown in blue represent the points at which $\tilde{\mathbf{s}}$ has constant ℓ_p-norm. Since the norms must be minimized, imagine the balls being inflated until they first touch the lines of possible solutions shown in red. For $0 < p \leq 1$, the ℓ_p-norms favor sparse solutions that lie on the axes. By comparison, minimizing the common Euclidean norm $(p = 2)$ is not useful since it does not obtain a solution which is necessarily sparse.

Conclusions

In this article, we learned how CS exploits redundancy to reduce sampling rates, and we saw under which conditions the original signal is preserved despite the system being underdetermined. We observed how universal sampling schemes requiring the existence but not knowledge of a sparsifying basis may be developed by using random sensing matrices. Finally, we saw how an ℓ_0-minimization algorithm can recover a signal with knowledge of the sparsifying basis, and considered the rationale behind ℓ_p-minimization algorithms for $0 < p \leq 1$.

Acknowledgments

I would like to thank Prof. Warren du Plessis, Prof. Pieter Jacobs, Marius-Corné Meijer, Wade Smith, and Winston Jimu for many insightful discussions. Sincere thanks are also given to the anonymous reviewers for their input, which helped better explain certain concepts.

Author

Heinrich E.A. Laue (laueheinrich@gmail.com) is a postgraduate student with the Department of Electrical, Electronic, and Computer Engineering at the University of Pretoria, South Africa. His research focuses on developing compressive feed networks for antenna arrays.

References

[1] R. G. Baraniuk, "Compressive sensing [Lecture Notes]" *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, July 2007.

[2] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[3] M. Fornasier and H. Rauhut, "Compressive sensing," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. New York: Springer New York, 2011, pp. 187–228.

[4] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. E. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[5] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, June 2010.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Resonance Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[7] A. Massa, P. Rocca, and G. Oliveri, "Compressive sensing in electromagnetics—A review," *IEEE Antennas Propag. Mag.*, vol. 57, no. 1, pp. 224–238, Feb. 2015.

[8] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[9] S. Foucart, "A note on guaranteed sparse recovery via ℓ_1-minimization," *Appl. Comput. Harmon. Anal.*, vol. 29, no. 1, pp. 97–103, July 2010.

[10] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, May 2015.

Ljubiša Stanković, Miloš Daković, and Ervin Sejdić

Vertex-Frequency Analysis: A Way to Localize Graph Spectral Components

Currently, brain and social networks are examples of new data types that are massively acquired and disseminated [1]. These networks typically consist of vertices (nodes) and edges (connections between nodes). Usually, information is conveyed through the strength of connection among nodes, but in recent years, it has been discovered that valuable information may also be conveyed in signals that occur on each vertex. However, traditional signal processing often does not offer reliable tools and algorithms to analyze such new data types. This is especially true for cases where networks (e.g., the strength of connections), or signals on vertices, have properties that change over the network.

This lecture note presents a new method to analyze changes in signals on graphs. This method, called the *vertex-frequency analysis*, relies on Laplacian matrices to establish connections between vertex changes and spectral content [2]–[5]. Specifically, this lecture note aims to connect concepts from frequency and time-frequency analyses (e.g., [6] and [7]) to the spectral analysis of graph signals. Graph signal processing is a major research area, however, we still lack understanding of how to relate graph signal processing concepts to concepts from traditional signal processing.

Digital Object Identifier 10.1109/MSP.2017.2696572
Date of publication: 11 July 2017

Relevance

The vertex-frequency analysis presented here is a valuable tool that can be used to analyze vertex-varying changes in networks (graphs) such as brain networks (e.g., brain changes during consecutive swallows [8]), changes in social interactions in a large group of people, or to understand traffic patterns during rush hour in major metropolitan areas. Theoretically, it connects principles of the Fourier analysis and eigenvalue decomposition from undergraduate courses, to more advanced topics such as time-frequency representations typically taught at a graduate level.

Prerequisites

The prerequisites for understanding this lecture note are linear algebra and an understanding of basic signal processing concepts.

Problem statement and solution

Problem statement

A graph consists of vertices and edges. If we denote the weights of graph edges connecting vertices n and m as w_{nm}, then the graph Laplacian operator is defined by

$$\mathbf{L} = \mathbf{D} - \mathbf{W},$$

where the matrix $\mathbf{W}$ elements are weighting coefficients w_{nm} and $\mathbf{D}$ is a diagonal matrix with elements $d_n = \sum_{m=1}^{N} w_{nm}$. An example of such a graph is shown in Figure 1(a).

Consider a signal $\mathbf{x}$ whose samples are $x(n)$, as shown in Figure 1(c), and these samples are assigned to (sensed at) the graph vertices as shown in Figure 1(b). The Laplacian operator applied to a signals on the graph is equal to $\mathbf{Lx}$, with elements $\mathcal{L}_x(n) = \sum_m w_{nm}(x(n) - x(m))$.

The spectral representation of a discrete-time signal $x(n)$ on the graph is defined as its expansion onto the set of eigenvectors (discrete-time basis eigenfunctions) of the Laplacian. To accomplish this expansion, the Laplacian $\mathbf{L}$ is decomposed as

$$\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^T,$$

where $\mathbf{U}$ is a matrix of the Laplacian eigenvectors $\mathbf{u}_k$ and Λ is a diagonal matrix of its eigenvalues λ_k.

The spectrum $X(\lambda_k)$ of signal $\mathbf{x}$ on a graph is calculated as its projection onto the corresponding eigenvectors $\mathbf{u}_k$ of the Laplacian:

$$X(\lambda_k) = \mathbf{u}_k^T\mathbf{x},$$

or in vector notation $\mathbf{X} = \mathbf{U}^T\mathbf{x}$.

Since the eigenvectors are orthogonal, the signal reconstruction is defined as $\mathbf{x} = \sum_{k=1}^{N} X(\lambda_k)\mathbf{u}_k = \mathbf{UX}$.

Spectral decomposition of a graph signal is illustrated in Figure 1(d). This spectrum contains three components corresponding to the constant component at $\lambda_1 = 0$, a low-frequency component at $\lambda_2 = 0.6934$, and a high-frequency component at $\lambda_6 = 2.4644$. We can split the signal into, for example, its low-frequency part by summing over $k = 1, 2$ and the high-frequency part by using $k = 6$ as $X(\lambda_6)\mathbf{u}_6$.

The meaning of weighting coefficients w_{nm} in a graph is highly dependent on the application, especially as the graph Laplacian is defined by these coefficients, and the Laplacian operator then defines the set of basis functions for signal expansion. For example, classical Fourier analysis can be obtained by considering the second-order derivative estimation (Laplace operator); see "Fourier Analysis and Laplacian." The Laplacian operator is also known as the *Kirchhoff matrix* in electrical circuit theory; see "Laplacian, Kirchhoff, and Ohm's Laws on an Electric Circuit Graph." In image processing, the coefficients w_{nm} may be proportional to the similarities of adjacent image pixels. Similarly, graphs are widely used in neuroscience, and edge coefficients are used to describe the strength of interactions among brain regions. In the case of a graph signal corresponding to a Euclidian network, the coefficient values are related to the vertex distances. A common way to define the coefficients in such networks is $w_{nm} = \exp(-r_{nm}/\tau)$ for $w_{nm} > \kappa$ and $w_{nm} = 0$ elsewhere, where r_{nm} is the Euclidian distance between vertices n and m, and τ and κ are constants. This approach is used in Figure 1(a).

The presented spectral analysis of signals on graphs provides a way to process signals in the graph spectral domain, that is, to implement signal processing techniques such as filtering, denoising, or to reconstruct missing signal values at some vertices if the graph signal spectrum is sparse.

Similar to Fourier domain analysis in traditional signal processing, the considered spectral analysis of signals on graphs has its limitations. For example, let us consider the graph shown in Figure 2(a) and two signals on this graph, presented in Figure 2(b) and (c). While these two signals $x_1(n)$ and $x_2(n)$ are obviously different, their spectral representations on this graph $|X_1(\lambda_k)|$ and $|X_2(\lambda_k)|$ are almost the same as depicted in Figure 2(d) and (e). Hence, it would be very difficult to implement any machine-learning schema that would be able to differentiate these two cases in the spectral domain. Therefore, an analysis is needed, similar to the time-frequency analysis in traditional signal processing, that is able to provide localized vertex information about the spectrum.

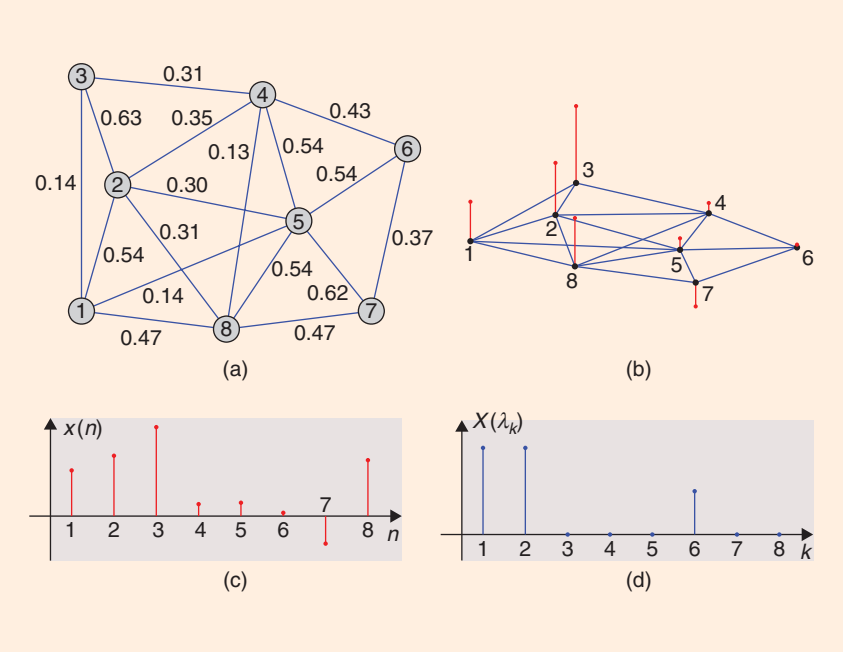


FIGURE 1. Sample graphs: (a) a graph with weighted edges; (b) a sample graph with signal values on each vertex; (c) a signal, $x(n)$; and (d) the spectrum of $x(n)$.

Solution

Localizing a spectral content around each vertex n can be achieved via vertex-frequency analysis. This analysis is an extension of the traditional time-frequency analysis to graph signals. As in the classical time-frequency analysis, a spectral transformation of a signal localized around the considered vertex n yields the basic formulation of the vertex-frequency analysis. This spectral transformation is typically achieved using a localization window. While different approaches exist, we will present two of them, one based on shifting a window in the vertex-frequency domain, and the other based on a vertex neighborhood analysis.

Convolution-based definition

To define a localized spectrum, let us consider two signals $x(n)$ and $h(n)$ on a graph with the corresponding Laplacian $\mathbf{L}$, whose eigenvalues and eigenvectors are λ_k and $u_k(n)$, respectively, while signal spectra are given by $X(\lambda_k)$ and $H(\lambda_k)$. Here, the signal $h(n)$ is used to localize the spectral characteristics of $x(n)$. For these two graph signals Parseval's theorem is given by $\Sigma_{n=1}^{N} x(n)h(n) = \Sigma_{k=1}^{N} X(\lambda_k)H(\lambda_k)$.

The shift of a signal on a graph cannot be extended in a direct way from the traditional signal processing theory. Hence, a generalized convolution operator on graphs is defined under the assumption that the spectrum of convolution $y(n) = x(n) * h(n)$ on a graph is equal to the product of signal spectra $Y(\lambda_k) = X(\lambda_k)H(\lambda_k)$. The convolution is then equal to the inverse transform of $Y(\lambda_k)$,

$$y(n) = x(n) * h(n) = \sum_{k=1}^{N} X(\lambda_k)H(\lambda_k)u_k(n).$$

Fourier Analysis and Laplacian

Fourier analysis uses the idea that a signal, $x(t)$, can be expanded in terms of orthogonal basis functions $\cos(2\pi ft)$ and $\sin(2\pi ft)$. In other words, the resulting Fourier representation is a projection (scalar product) of the signal onto sinusoidal basis functions.

Interestingly enough, the Fourier expansion can be also considered from the Laplacian (Laplace differential operator), $\mathcal{L}\{x(t)\} = -d^2x(t)/dt^2$, point of view. The Laplacian eigenfunctions $u(t)$ are the solutions of $\mathcal{L}\{u(t)\} = \lambda u(t)$. We can easily conclude that $\cos(2\pi ft)$ and $\sin(2\pi ft)$ are the eigenfunctions of the Laplacian with the eigenvalues $\lambda = (2\pi f)^2$. Therefore, the Fourier analysis can be defined as an expansion of the signal $x(t)$ onto the set of eigenfunctions of the Laplacian operator.

In the discrete-time domain the Laplacian can be defined using a symmetric second-order difference operator

$$\begin{aligned}\mathcal{L}\{x(n)\} &= -x(n-1) + 2x(n) - x(n+1)\\ &= (x(n) - x(n-1)) + (x(n) - x(n+1))\\ &= \sum_{m\in\{n-1,n+1\}} w_{nm}(x(n) - x(m)).\end{aligned}$$

The Laplacian is a matrix $\mathbf{L}$ that can be used to transform a discrete-time signal $x(n)$ into its second-order difference. Eigenvectors of this Laplacian are the discrete-time sine and cosine functions.

A graph corresponding to circular form of this Laplacian, sample signal $x(n)$, and corresponding spectrum $X(\lambda_k)$, for $N = 8$, are shown in Figure S1. Signal values are assigned to graph vertices, and the resulting spectrum $X(\lambda_k)$ is obtained by decomposing $x(n)$ onto the Laplacian eigenvectors. The Laplacian matrix (in a circular form) is

$$\mathbf{L} = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & -1\\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1\\ -1 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

The Laplacian operator maintains the relationship among vertices (signal samples) regardless of their ordering. Even if vertices (signal samples) are arbitrarily reordered, the eigenvectors of the Laplacian produce the same spectrum.

If the signal vector $\mathbf{x}$ is a stacked-column representation of a two-dimensional $N \times N$ image, then with the summation for $m \in \{n-N, n-1, n+1, n+N\}$, when $w_{nm} = 1$ in $\mathcal{L}\{x(n)\}$, the two-dimensional Fourier analysis can be defined.

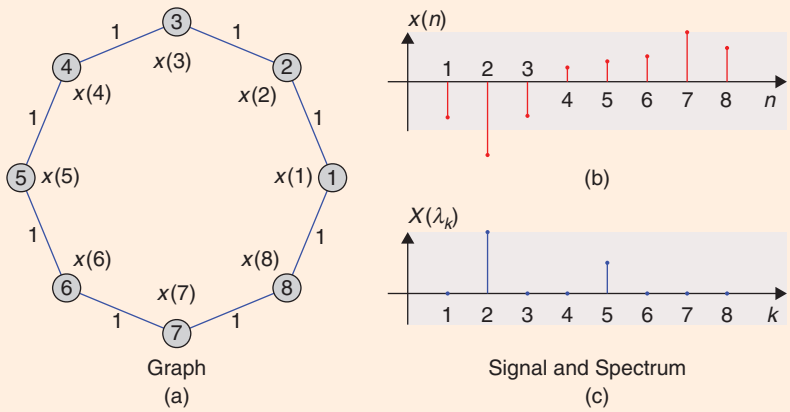


FIGURE S1. A (a) graph for Fourier signal analysis, (b) signal on a graph, and (c) the corresponding spectrum.

This is the definition of the generalized convolution operator of two signals on a graph [3].

Convolution can be used to define the shift on a graph as $h(m-n) = h(m) * \delta_n(m) = \sum_{k=1}^{N} H(\lambda_k)u_k(n)u_k(m)$ where $\delta_n(m)$ is the delta function at the nth vertex. Its spectrum $\Delta_n(\lambda_k)$ is equal to the nth sample of the kth eigenfunction, since $\Delta_n(\lambda_k) = \sum_{m=1}^{N}\delta_n(m)u_k(m) = u_k(n)$.

The localized vertex spectrum (LVS) on a graph can be calculated as the spectrum of a graph function $x(n)$ multiplied by a shifted window $h(n-m)$

$$\begin{aligned}\mathrm{LVS}_x(n,\lambda_k) &= \sum_{m=1}^{N} x(m)h(m-n)u_k(m)\\ &= \sum_{m=1}^{N} x(m)h_{n,k}(m),\end{aligned}$$

where the localized version of the window on vertex and frequency axes is denoted by

$$h_{n,k}(m) = \sum_{l=1}^{N}[H(\lambda_l)u_l(n)u_l(m)]u_k(m),$$

where we can use, for example, $H(\lambda_k) = C\exp(-\lambda_k\tau)$.

The inverse formula is then a sum of $\mathrm{LVS}_x(n,\lambda_k)$, multiplied by

Laplacian, Kirchhoff, and Ohm's Laws on an Electric Circuit Graph

Kirchhoff and Ohm's laws on a graph

The Laplacian can be considered from the basic electric circuit theory point of view (Kirchhoff matrix). Let us assume that a graph represents an electric circuit. Then, the signal values $x(n)$ represent node voltages at the corresponding circuit vertices $x(n) = v(n)$. The weight coefficients $w_{nm} = 1/R_{nm}$ represent conductance (reciprocal resistance R_{nm}) values in the edges connecting vertices n and m, for vertices that are not connected by an edge $w_{nm} = 0$. The value of the current in the edge from vertex n to m is given by $i_{nm} = (v(n) - v(m))/R_{nm} = w_{nm}(x(n) - x(m))$. The sum of all currents going into a vertex n must be 0. In general, the external current source connected to vertex n is equal to the sum of all currents going from this vertex, $i_G(n) = \Sigma_i w_{nm}(x(n) - x(m)) = d_n x(n) - \Sigma_m w_{nm} x(m)$, where $d_n = \Sigma_m w_{nm}$.

The matrix form of the voltage to current relation is $\mathbf{Lx} = \mathbf{i}_G$, where $\mathbf{L}$ is the Laplacian of the graph (circuit). The node (vertex) voltage vector $\mathbf{v} = \mathbf{x}$ is determined (up to the constant referent voltage) from the vector of external currents $\mathbf{i}_G$ via a system of linear equations.

As in the Fourier analysis, the solution of this system can be simplified using the spectral decomposition of the current and the node voltage vectors onto the set of eigenvectors of the Laplacian. Starting with $\mathbf{Lv} = \mathbf{U}\Lambda\mathbf{U}^T\mathbf{v} = \mathbf{i}_G$ and understanding that $\mathbf{U}^T\mathbf{U}$ is a unitary matrix, we obtain $\Lambda\mathbf{U}^T\mathbf{v} = \mathbf{U}^T\mathbf{i}_G$. This represents an Ohm's law analog on a graph, $\lambda_k V(\lambda_k) = I(\lambda_k)$, where $V(\lambda_k) = \mathbf{u}_k^T\mathbf{v}$ and $I(\lambda_k) = \mathbf{u}_k^T\mathbf{i}_G$ are the spectral coefficients of the vertex voltage $x(n) = v(n)$, and the external vertex current $i_G(n)$ on the graph.

A similar analysis can be performed for a heat transfer flux, with edge weights representing heat transfer coefficients in an appropriate thermodynamics problem definition.

the shifted and modulated windows $h_{n,k}(m)$,

$$x(n) = \frac{1}{\sum_{k=1}^{N} |H(\lambda_k)|^2 |u_k(n)|^2} \times \sum_{i=1}^{N}\sum_{k=1}^{N} LVS_x(i,\lambda_k) h_{i,k}(n).$$

Note that the outlined approach [3] can be computationally expensive, and a fast implementation algorithm is proposed in [9].

Definition based on vertex neighborhood

To obtain a localized spectrum of a graph signal, we can utilize localization functions (windows) corresponding to window functions in classical signal processing. As in classical signal processing, a window function should be narrow enough to provide good localization of the spectral components but wide enough to produce high resolution of such components. In other words, the window should contain the considered signal sample and some neighboring vertex samples. That is, the window is defined by a set of vertices that contain the current vertex n and all vertices that are close to the nth vertex.

There are several ways to define the local neighborhood for a vertex. For example, we can consider that two vertices are close if there is an edge between them, or if there is a path with its length (number of edges) smaller than an assumed threshold. Edge weights could also be taken into consideration to decide whether two nodes are close enough or not.

Commonly, the edge weights are given by $w_{nm} = \exp(-r_{nm}/\tau)$ for $r_{nm} < \kappa$, and $w_{nm} = 0$ otherwise. Here, r_{nm} denotes a distance between vertices, while τ and κ are constants. If we consider two arbitrary vertices n and m on a graph, then the path weight between these two vertices can be defined as the product of all edge weights that are included in the considered path: $p_{nm} = w_{nk_1}w_{k_1k_2}\cdots w_{k_bm}$. If there is more than one path between nth and mth vertices, the shortest path (with the highest p_{nm} value) is considered. It can also be stated that the vertex m belongs to the local neighborhood of the nth vertex if $p_{nm} \geq h_T$, where h_T is a threshold defining the window size.

The simplest window has a value of $h_n(m) = 1$ for all vertices m that belong to the window centered at the nth vertex, and $h_n(m) = 0$ otherwise. It is analogous to a rectangular window in traditional signal processing. When $h_n(m) = 1$ for each m, the standard spectrum is obtained on a graph. We can define window function values $h_n(m)$ based on the distances p_{nm} that will attenuate farther vertex samples. Now we can define the signal localized around the nth vertex as $x_h(m) = x(m)h_n(m)$. The corresponding local spectrum is then defined as:

$$\mathrm{LVS}_x(n,\lambda_k) = \sum_{m=1}^{N} x(m)h_n(m)u_k(m),$$

and the inverse definition follows from the inverse spectrum relation with additional summation over all vertices:

$$x(n) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{N} \mathrm{LVS}_x(i,\lambda_k)u_k(n)}{\sum_{i=1}^{N} h_i(n)}.$$

Note that for the windowed signal $x(m)h_n(m)$, only $M \leq N$ samples are nonzero, meaning that it can be considered as a zero padded signal. To reconstruct this signal, we only need M spectral coefficients $\mathrm{LVS}_x(n,\lambda_k)$ for M different values of λ_k. The remaining coefficients can be calculated from the system of equations obtained by using the fact that $x(m)h_n(m) = 0$

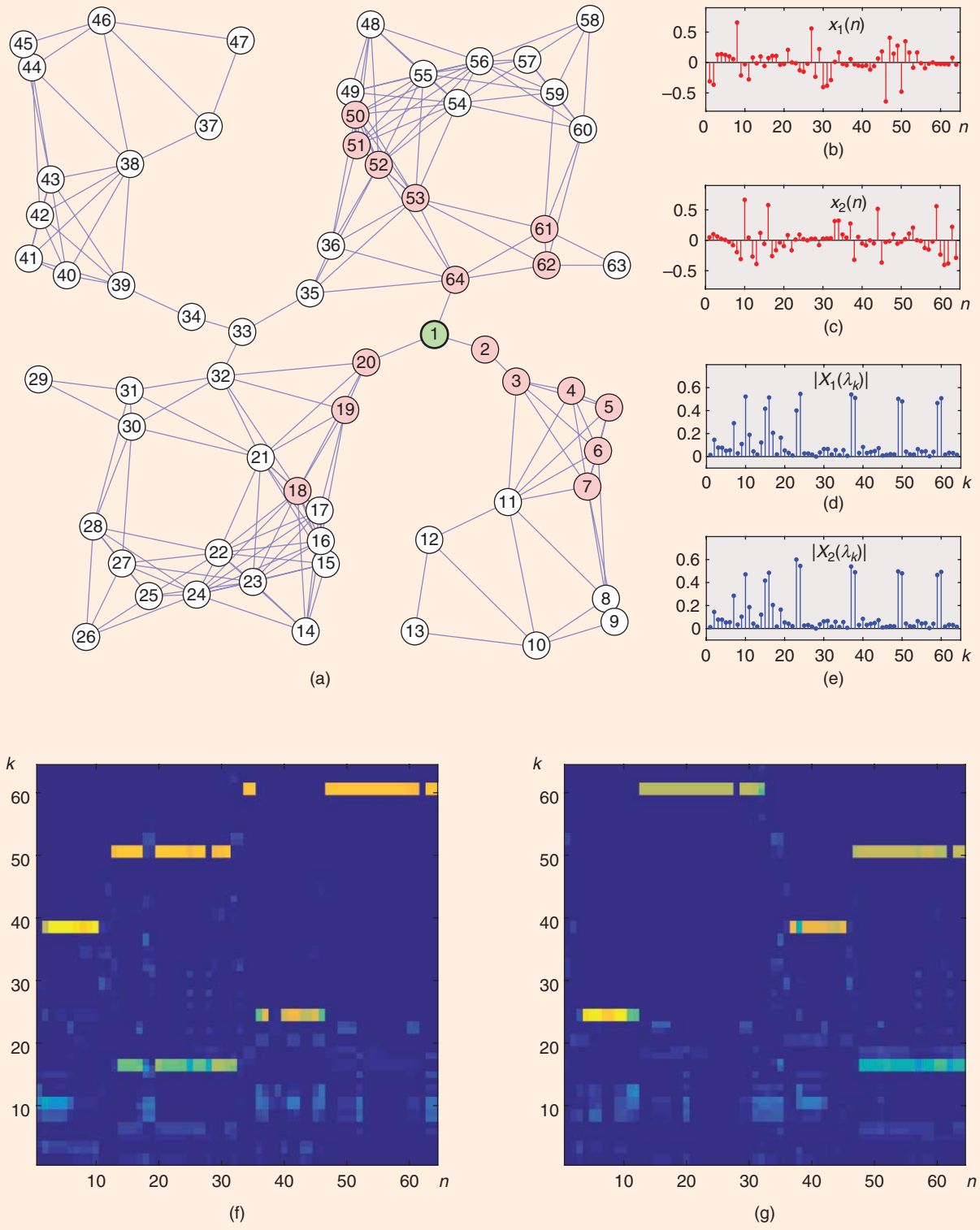


FIGURE 2. Graph signal processing in action: (a) a graph with shaded nodes belonging to a localization window centered at the first vertex (the green-shaded vertex); (b) a sample signal, $x_1(n)$; (c) a sample signal, $x_2(n)$; (d) the graph spectrum, $|X_1(\lambda_k)|$, of $x_1(n)$; (e) the graph spectrum, $|X_2(\lambda_k)|$, of $x_2(n)$; (f) a vertex-frequency representation of $x_1(n)$; and (g) a vertex-frequency representation of $x_2(n)$.

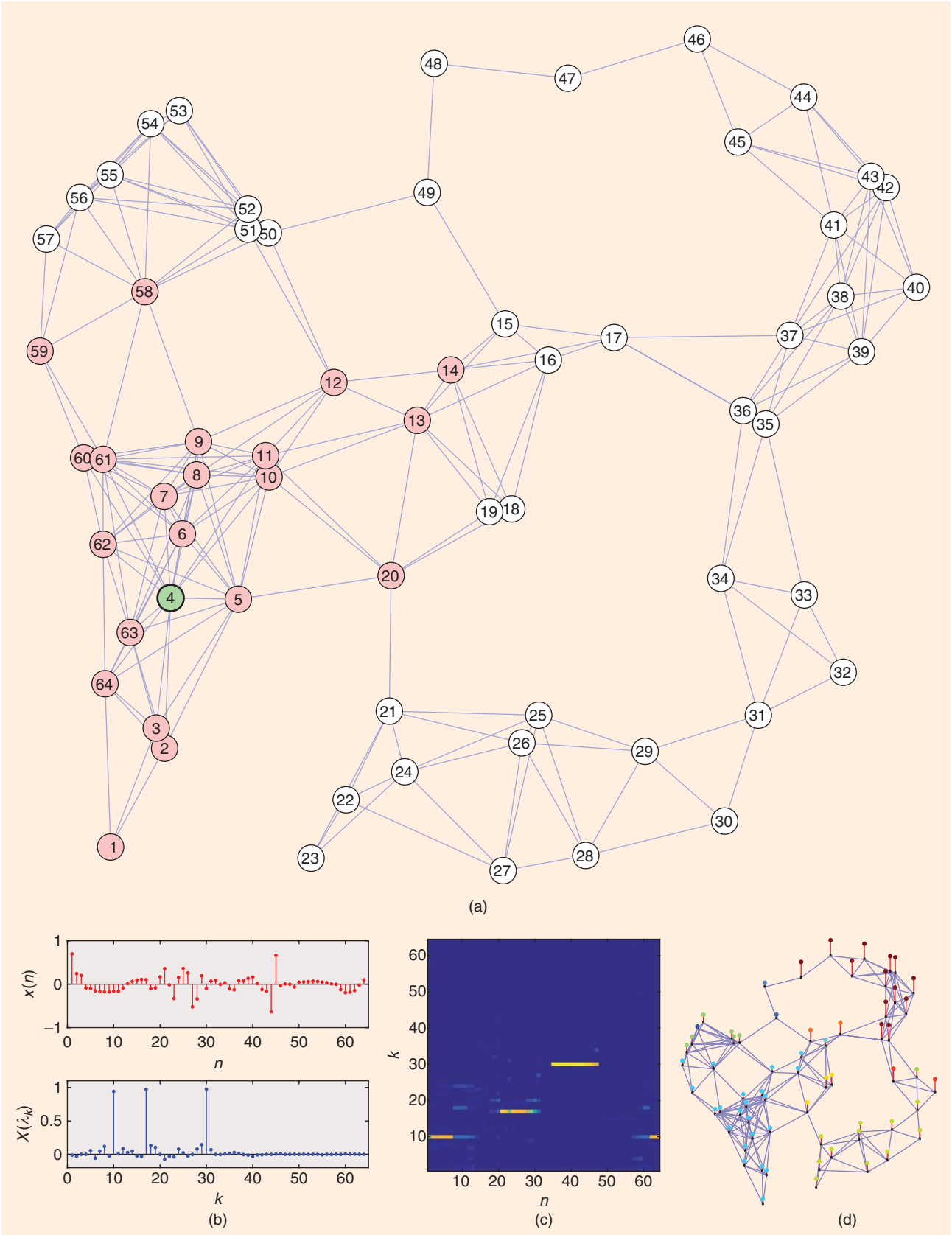


FIGURE 3. The vertex-frequency analysis of a graph signal: (a) a graph and a window centered at the fourth vertex (the vertices included in the window are shaded); (b) a graph signal and its spectrum; (c) a vertex-frequency representation of the graph signal; and (d) an instantaneous (vertex-based) frequency representation.

outside the window support. It produces a system of $N-M$ linear equations $\Sigma_{k=1}^{N}\mathrm{LVS}_x(n,\lambda_k)u_k(m)=0$ for vertices outside the window support. This system provides conditions for the spectral coefficients "interpolation" using M-calculated values $\mathrm{LVS}_x(n,\lambda_k)$.

To visualize the local spectral content, we should order vertices, i.e., find the Hamiltonian path in the corresponding graph. This ordering is not unique, and a possible way for ordering is to keep in mind that neighboring vertices have the highest possible edge weights.

Numerical examples

Example 1

Let us first consider a vertex-frequency analysis of two graph signals shown in Figure 2(b) and (c). For each vertex n we can define a window $h_n(m)$ and calculate the local spectrum $\mathrm{LVS}_x(n,\lambda_k)$ of the windowed signal. A localized support for a window, centered at the first vertex, is presented in Figure 2(a) with red-shaded vertices corresponding to the window support. In this way, we obtain a two-dimensional representation $\mathrm{LVS}_x(n,k)$ of the analyzed signal presented, as a function of the vertex and eigenvalue index, in Figure 2(f) and (g). We can see that different signals, having almost the same spectrum on graph, have different vertex-frequency representations.

Example 2

Consider a signal, $x(n)$, defined on a graph with $N=64$ vertices as presented in Figure 3(a). Let's assume that the signal values $x(n)$ are defined with Laplacian eigenvectors $u_k(n)$ as: $x(n)=u_{17}(n)$ for $17\le n\le 32$, $x(n)=u_{30}(n)$ for $33\le n\le 48$, and $x(n)=u_{10}(n)$ otherwise.

Signal samples and its spectrum are given in Figure 3(b). The signal spectrum clearly depicts peaks at $k=10$, $k=17$, and $k=30$. Small spectrum values (side lobes) around these eigenvalues exist since the components are not complete over all vertices. The vertex-frequency analysis of this signal is performed using the localization window $h_n(m)$. A localization area for the $h_4(m)$ window, centered at the fourth vertex, is shown in Figure 3(a) (red-shaded vertices). The local spectrum $\mathrm{LVS}_x(n,\lambda_k)$ of the windowed signal is calculated and presented in Figure 3(c). From this representation, we can see localized signal components at "frequencies" $k=10$, $k=17$, and $k=30$.

Finally, an instantaneous frequency representation is provided in Figure 3(d). It should be noted that the "instantaneous frequency" definition for graphs is different from such a definition in traditional signal processing, where instantaneous frequency is defined as a signal's phase derivative with respect to time. Here, we determine the "frequency" (eigenvalue) index at each vertex, and this "frequency" index represents the kth index for which the spectrum reaches maximum at that particular vertex. Next, we can plot vertical lines with their lengths and colors proportional to the position (frequency) of the spectral maximum for each vertex as depicted in Figure 3(d). This essentially yields a vertex-based instantaneous frequency representation depicting the localization of signal components on graph vertices.

What we have learned

Graph signal processing is a new field that compliments traditional signal processing. While traditional signal processing techniques for the analysis of time-varying signals are well established, the corresponding graph signal processing equivalent approaches are in their infancy. In this lecture note, we presented novel algorithms for the analysis of vertex-varying graph signals. We expect that the considered technique will find its many uses in neuroscience, social sciences, and genome processing, as graphs (networks) in those applications tend to be "nonstationary," and current analytical tools widely ignore this fact. Hence, the vertex-frequency analysis is of paramount importance for such applications.

Authors

Ljubiša Stanković (ljubisa@ac.me) is a professor at the University of Montenegro. His research interests include digital signal processing and time-frequency analysis. He is a Fellow of the IEEE.

Miloš Daković (milos@ac.me) is a professor at the University of Montenegro. His research interests include signal processing and time-frequency analysis. He is a Member of the IEEE.

Ervin Sejdić (esejdic@ieee.org) is an associate professor at the University of Pittsburgh, Pennsylvania. His research interests include biomedical signal processing, rehabilitation engineering, neuroscience, swallowing, and gait. He received the Presidential Early Career Award for Scientists and Engineers in 2016 and the National Science Foundation CAREER Award in 2017. He is a Senior Member of the IEEE.

References

[1] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sept. 2014.

[2] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[3] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 260– 291, 2016.

[4] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[5] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE. Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, June 2014.

[6] L. Stanković, M. Daković, and T. Thayaparan, *Time-Frequency Signal Analysis with Applications*. Boston, MA: Artech House, 2013.

[7] B. Boashash, Ed, *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. New York: Academic, 2015.

[8] I. Jestrović, J. L. Coyle, and E. Sejdić, "Differences in brain networks during consecutive swallows detected using an optimized vertex–frequency algorithm," *Neuroscience*, vol. 344, pp. 113–123, Mar. 2017.

[9] I. Jestrović, J. L. Coyle, and E. Sejdić, "A fast algorithm for vertex-frequency representations of signals on graphs," *Signal Processing*, vol. 131, pp. 483–491, Feb 2017.

SP

Richard Lyons

Digital Envelope Detection: The Good, the Bad, and the Ugly

During a recent consulting job to analyze acoustic telemetry signals transmitted by a deep-sea drill pipe, I was forced to investigate a process called *digital envelope detection*. This process is used to estimate the instantaneous magnitude of a zero-mean fluctuating-amplitude digital signal. While much tutorial information regarding envelope detection is available, that information is spread out over a number of communications textbooks and many websites. The purpose of this article is to collect and describe various digital envelope detection methods in one concise and consistent lesson.

Envelope detection is used in a wide variety of signal processing applications, where we want to detect the presence of a narrowband signal or estimate the instantaneous energy of a signal. Such applications include amplitude modulation (AM) radio demodulation, automatic gain control, modulated optical signal detection, medical blood pressure evaluation, and ultrasound signal analysis, just to name a few.

The only prerequisites for understanding this article are knowledge of the fundamental fluctuating-amplitude nature of discrete time signals and familiarity with simple digital signal processing (DSP) signal flow, i.e., block diagrams.

Problem of envelope detection

The problem solved by envelope detection is to acquire a fluctuating-amplitude sinusoidal discrete signal where the positive-amplitude fluctuations, i.e., the sinusoid's envelope, contain some sort of desired information and to extract that information. An example of such a sinusoid is the amplitude modulated radio-frequency (RF) signal shown in Figure 1(a). The dashed curve in that figure represents the RF signal's $m(n)$ envelope, and it is the goal of envelope detection to extract and make available that envelope signal as shown in Figure 1(b).

An important note: although the signal waveforms in Figure 1 appear to be continuous, keep in mind that they are indeed discrete-time quantized numerical sequences, i.e., digital signals. Furthermore, the blue curve in Figure 1(a) represents a fluctuating-amplitude $\cos(2\pi f_c n t_s)$ discrete sinusoidal sequence whose frequency is f_c Hz, and t_s is the reciprocal of the sequence's f_s sample rate.

Possible solutions

DSP practitioners have developed a remarkably wide variety of different methods to perform envelope detection over the past few decades. Here I illustrate their cleverness by describing today's most common envelope detection techniques.

Asynchronous half-wave envelope detection

Figure 2(a) is a digital version of the popular diode envelope detector used in the analog world for AM radio demodulation. Here, the thresholding operation sets all the negative-valued samples in the modulated RF input sequence to zero, a process we rightfully call *half-wave rectification*. This simple envelope detector is called *asynchronous* because it does not generate a constant-amplitude copy of the incoming $\cos(2\pi f_c n t_s)$ RF sinusoid, as do some of the other detectors that will be discussed in this article.

Due to the harmonics, i.e., multiples of the incoming f_c carrier frequency, generated by the nonlinear half-wave rectification in Figure 2(a), and possible spectral aliasing depending on the system's f_s sample rate, careful spectrum

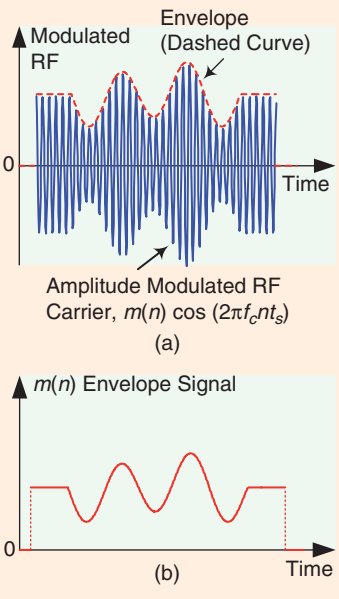


FIGURE 1. (a) A fluctuating-amplitude RF signal and (b) its $m(n)$ envelope signal.

Digital Object Identifier 10.1109/MSP.2017.2690438
Date of publication: 11 July 2017

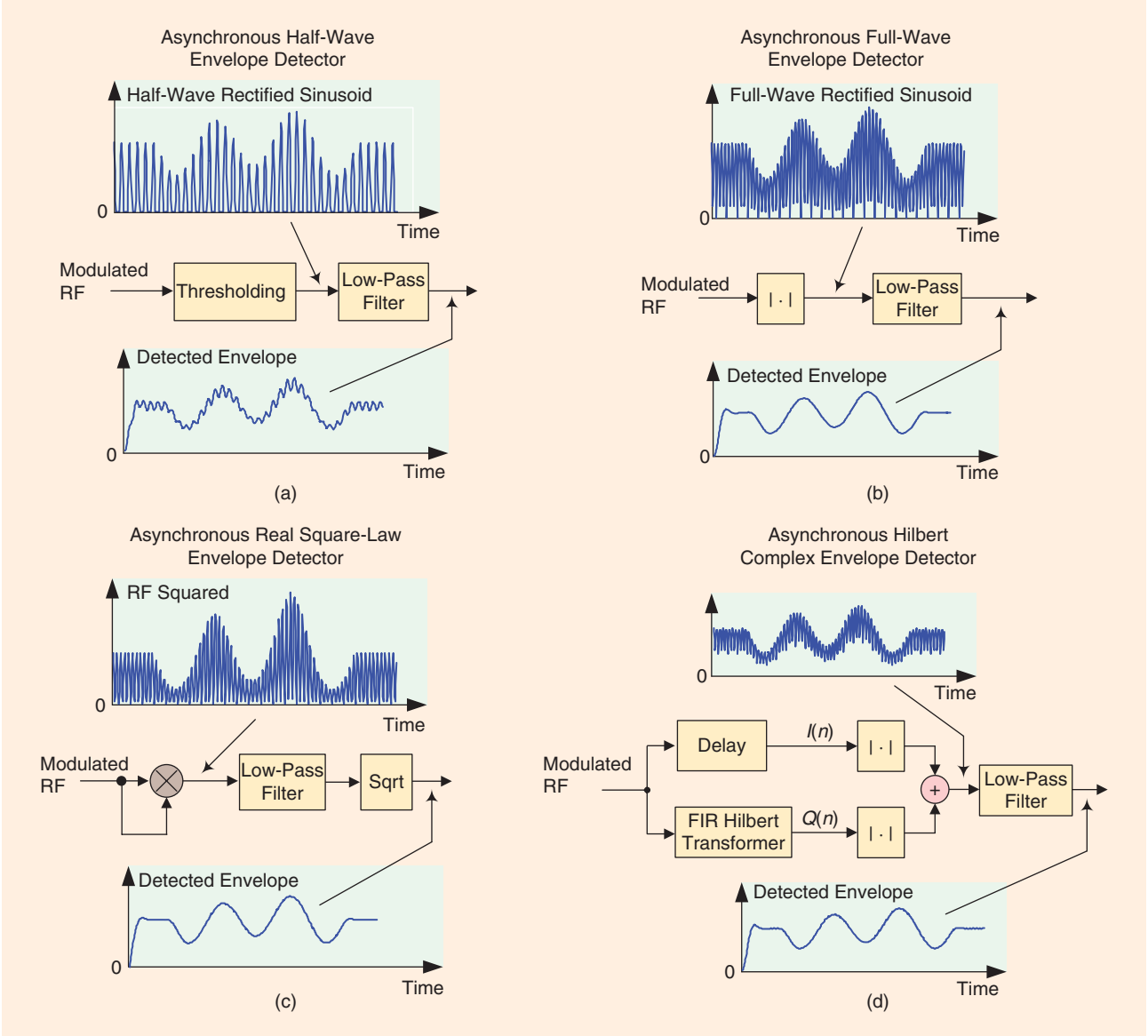


FIGURE 2. Four methods of asynchronous envelope detector are shown in (a)–(d).

analysis of the half-wave rectified sinusoid is necessary to help you determine the appropriate cutoff frequency of the digital low-pass filter.

Asynchronous full-wave envelope detection

We can reduce the high-frequency noise riding on Figure 2(a) detector's output by performing full-wave rectification as shown in Figure 2(b) [1]. In Figure 2(b), the $|\cdot|$ symbol means the computation of the absolute value.

Here, the lowest-frequency spectral harmonic at the filter's input is $2f_c$ Hz, thus that harmonic is more thoroughly attenuated at the Figure 2(b) low-pass filter output compared to the f_c Hz amplitude fluctuations at the Figure 2(a) filter output.

Asynchronous real square-law envelope detection

Figure 2(c) is a digital version of a popular square-law analog envelope detector, sometimes called a *product detector* or a *root mean square* (*RMS*) *demodulator*. Here, the spectral harmonics at the low-pass filter's input are roughly the same as those in Figure 2(b). This envelope detection method is used in Analog Devices' AD8361 Detector integrated circuit. Reference [2] gives a mathematical description of the Figure 2(c) square-law envelope detector.

Asynchronous Hilbert complex envelope detection

Figure 2(d) shows a popular digital envelope detector that uses a finite impulse response (FIR) Hilbert transformer to compute an $I(n) + jQ(n)$ complex-valued version of the incoming signal. The Delay element, whose time delay is equal to the Hilbert transformer's group delay, measured in samples, is required to achieve discrete signal synchronization

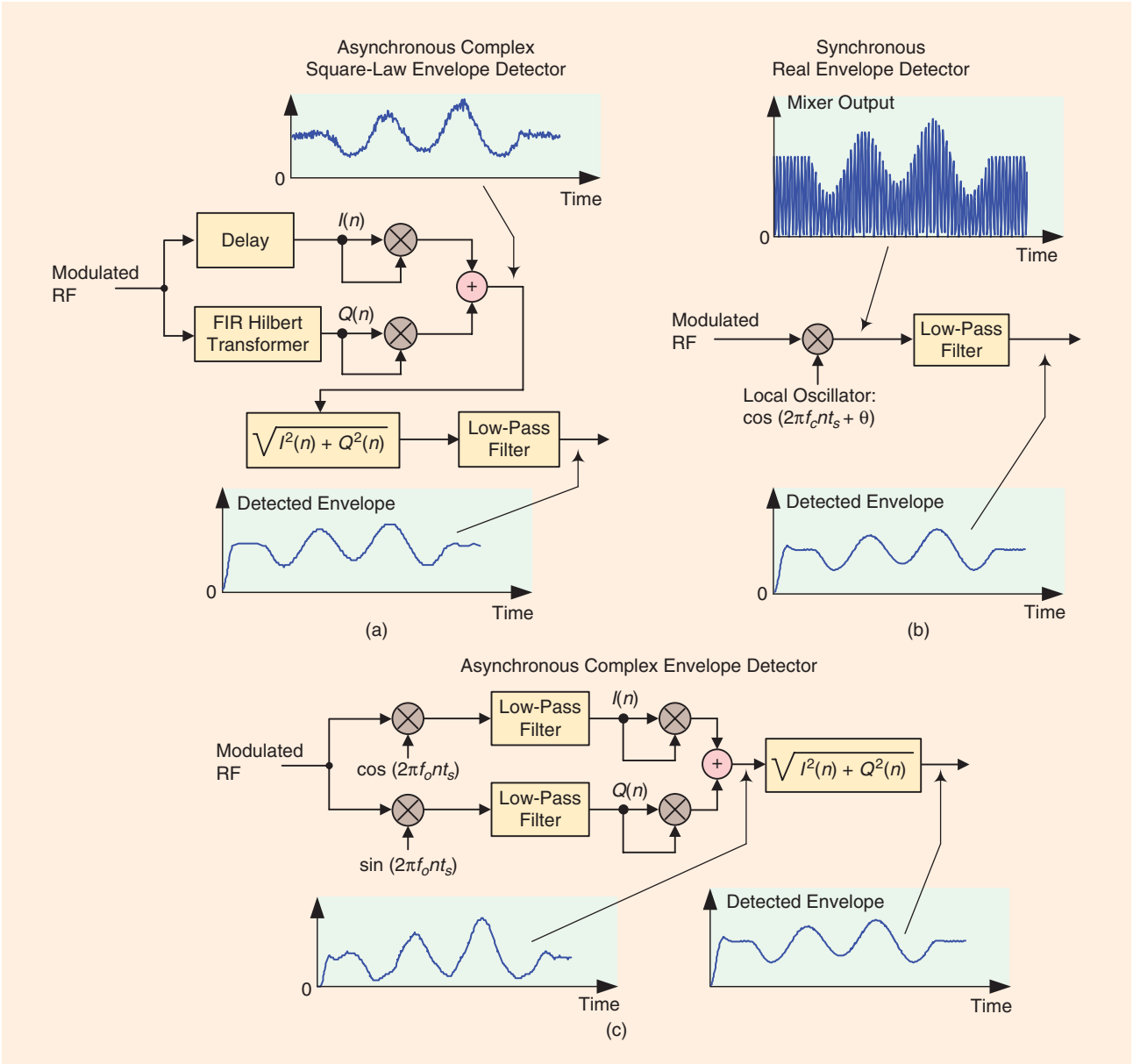


FIGURE 3. (a) and (c) show two asynchronous envelope detector methods while (b) shows one synchronous envelope detector method.

with the delayed output of the Hilbert transformer. This detector, using an exponential averaging low-pass filter, is described in more detail in [3].

Asynchronous complex square-law envelope detection

Another digital envelope detector that uses a Hilbert transformer is shown in Figure 3(a). Reference [4] touts this detector's advantage that no low-pass filtering is needed at the output of the square root operation. However, I have learned this only to be true for noise-free modulated RF input signals! In practical real-world applications, the low-pass filter shown in Figure 3(a) is necessary. Reference [2] presents a mathematical description of this complex square-law envelope detector.

Synchronous real envelope detection

Figure 3(b) shows an envelope detector that is called *synchronous* because the modulated RF input signal is multiplied by a local oscillator signal whose frequency is f_c Hz. This detector is sometimes called a *coherent envelope detector.*

The complicated part of this detector is that the f_c Hz carrier frequency of the received RF input signal must be regenerated, which is a process called *carrier recovery*, within the envelope detector to provide the local oscillator's $\cos(2\pi f_c n t_s + \theta)$ signal. A phase error, θ, between the received RF signal's carrier and the local oscillator will cause a low-pass filter output amplitude reduction because that output is proportional to $\cos(\theta)$. A small constant θ can be tolerated, but a fluctuating value for θ causes unacceptable output signal "fading." Furthermore, it is critical to

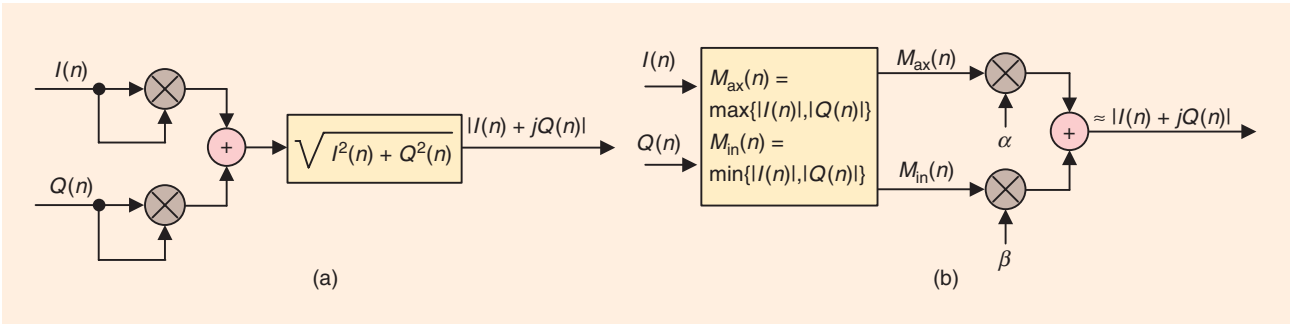


FIGURE 4. Complex sequence magnitude estimation: (a) the traditional method and (b) the alpha max plus beta min method.

ensure the local oscillator's frequency is within a few hertz of the input RF carrier's frequency.

It's worth noting that the envelope detector in Figure 3(b) is used in Texas Instruments' LM1596 and Analog Devices' AD630 Modulator/Demodulator integrated circuits. Reference [1] presents a mathematical description of this synchronous envelope detector.

Asynchronous complex envelope detection

Reference [5] proposes the Figure 3(c) envelope detector that computes a complex-valued version of the incoming signal. This detector is asynchronous in operation because the local oscillators' f_o frequency need not be equal to the f_c frequency of the incoming modulated RF. An f_o within 25% of f_c is acceptable so long as the low-pass filters sufficiently attenuate spectral energy near $|f_c + f_o|$ Hz.

Because the $I(n)$ and $Q(n)$ sequences are in quadrature, phase cancelation occurs in the adder such that only spectral energy in the vicinity of 0 Hz appears at the adder's output. As such, this detector is very tolerant of local oscillator frequency/phase drift just so long as the local cosine/sine oscillators remain in quadrature. This envelope detector is used in CML Microcircuits' CMX991 RF Quadrature Transceiver integrator circuit.

Envelope detection is used in a wide variety of signal processing applications, where we want to detect the presence of a narrowband signal or estimate the instantaneous energy of a signal.

The detector in Figure 3(c) is the final envelope detector we present in this article. Having said that, let's now discuss a few practical implementation issues regarding the aforementioned envelope detectors.

DC bias on the modulated RF

The described behavior of the aforementioned envelope detectors is based on the assumption that the dc bias, which is the average value, of the incoming modulated RF signal is small, i.e., less than 5% of the maximum value of the modulated RF. Happily, this restriction will always be satisfied if the modulated RF signal was generated by today's commercially available analog to digital converters.

Downsampling

I have encountered several websites that suggest digital downsampling, i.e., decimation, is appropriate at various stages within a given envelope detector. For example, [6] shows downsampling following the squaring operation in Figure 2(c) as well following the square root operation in Figure 3(a) prior to any low-pass filtering. Downsampling may be beneficial at various stages within an envelope detector, but you must ensure that any downsampling does not violate the Nyquist criterion. Failing to do so may force you to update your job resume!

Hilbert transformers

The Hilbert transformers in the Figures 2(d) and 3(a) envelope detectors need not be superhigh performance, such as a wide-band Hilbert transformer whose passband extends from nearly 0 Hz to nearly half the sample rate ($f_s/2$ Hz). The transformers' passbands need only be wide enough to include the spectral energy of the incoming modulated RF signal.

Complex magnitude estimation

You may have noticed that two of the previously mentioned envelope detectors require the computation of the magnitude of the complex sequence $I(n) + jQ(n)$, as shown in Figure 4(a). Square root operations are computationally expensive, so in real-time applications you might consider the computationally efficient magnitude estimation scheme shown in Figure 4(b).

Called the *alpha max plus beta min method*, the processing in Figure 4(b) replaces the troublesome square root operation with a few simple logical comparison operations. Variables α and β are constants, and for various values of those constants the magnitude estimation error can be as little as 1%. The error performance for various combinations of α and β can be found in [7].

Detector performance: The good, the bad, and the ugly

You'll notice that we've presented no statistical information on the signal-to-noise ratio (SNR) performance of the aforementioned envelope detectors. In fact, that was not our goal in this article. However, based on my MATLAB

Table 1. Envelope detector performances.

Highest output SNR	Asynchronous complex square-law [Figure 3(a)]
↓	Asynchronous Hilbert complex [Figure 2(d)]
↓	Asynchronous full-wave [Figure 2(b)]
↓	Asynchronous real square-law [Figure 2(c)]
↓	Asynchronous complex [Figure 3(c)]
↓	Synchronous real [Figure 3(b)]
Lowest output SNR	Asynchronous half-wave [Figure 2(a)]

software modeling of the various envelope detectors, with

- Sample rate: $f_s = 8,000$ Hz
- RF carrier frequency: 600 Hz
- Modulation: 60 Hz sine +30 Hz cosine wave
- Modulated RF signal SNR: +23 dB
- LP filter: third-order Butterworth infinite impulse response (≈240 Hz cutoff frequency)

I rank the detectors' performances (from best to worst) as shown in Table 1.

Summary

Various popular methods of envelope detection were listed and briefly described. Although computationally simple to implement, the Figure 2(a) detector should be avoided due to its high output noise behavior.

For moderate-performance applications, such as AGC or analyzing medical signals, the detectors in Figures 2(b) and (c) and 3(b) and (c) are appropriate choices. Note that, despite its computationally simple implementation, the Figure 2(b) detector performs quite well compared to the other detectors in this moderate-performance category.

For high-performance applications, such as in digital communications systems, the detectors in Figures 2(d) and 3(a) are the preferred choices. While their SNR performances are very similar, note that the Figure 2(d) detector requires far fewer arithmetic operations per output sample than the Figure 3(a) detector.

As a general rule, if you need an envelope detector in your signal processing application, I suggest you implement several of the aforementioned detectors to see which method is optimum for your input signals, your f_s data sample rate, and your data throughput requirements. To quote Forrest Gump, "And that's all I have to say about that."

Author

Richard Lyons (R.Lyons@ieee.org) is a consulting signal processing engineer. Winner of the IEEE 2012 Education Award, he is the author of *Understanding Digital Signal Processing 3/E* (Prentice Hall, 2010). He is the editor of, and contributor to, *Streamlining Digital Signal Processing, A Tricks of the Trade Guidebook* (IEEE Press/Wiley, 2007) and the coauthor of *The Essential Guide to Digital Signal Processing* book (Prentice Hall, 2014).

References

[1] C. Johnson, Jr., W. Sethares, and A. Klein, *Software Receiver Design*. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 82–84.

[2] S. A. Tretter. Amplitude modulation. [Online]. Available: http://www.ece.umd.edu/~tretter/commlab/c6713slides/ch5.pdf

[3] R. Lyons, *Understanding Digital Signal Processing*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 2011, pp. 786–784.

[4] D. Ciardullo. A fast envelope detector, Brookhaven Nat. Lab., AGS/AD/Tech. Note No. 386. [Online]. Available: http://www.agsrhichome.bnl.gov/AGS/Accel/Reports/Tech%20Notes/TN386.pdf

[5] M. E. Frerking, *Digital Signal Processing in Communications Systems*. London: Chapman & Hall, 1994, pp. 235–238.

[6] MathWorks, Inc. Envelope detection. [Online]. Available: https://www.mathworks.com/help/dsp/examples/envelope-detection.html?s_tid=gn_loc_drop

[7] C. Turner. Fast magnitude calculation. [Online]. Available: http://www.claysturner.com/dsp/FastMagnitude.pdf

SP

Magdy Bayoumi

It Really Was Lagniappe!

Highlights from ICASSP 2017 in New Orleans

In Cajun slang, spoken in southern Louisiana, *lagniappe* means *something extra*, and the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2017, held 5–9 March in New Orleans, really was lagniappe. This beautiful and historic city welcomed ICASSP attendees and the IEEE Signal Processing Society (SPS) with open arms and rich culture. New Orleans is the heart of great bayous; the melting pot of Cajun, Zydeco, and Creole cultures; the capital of jazz music; and the home of Mardi Gras.

ICASSP 2017 drew 2,231 attendees from about 50 countries. The technical program was hot and spicy (the Cajun style). It was a streamlined program of 2,518 very high-quality papers. All papers were rigorously reviewed. This outstanding technical program was attributed to the excellent contributions and hard work of the technical program committees, reviewers, and, most importantly, the authors. The program included a set of 16 special sessions focused on emerging and futuristic technologies and visions, which were very well attended. The tutorials were extremely popular and addressed the state-of-the-art topics and technologies, and they attracted very large crowds. The tutorial "Methods for Interpreting and Understanding Deep Neural Networks" drew more than 160 people!

One of the main highlights of the technical program was four keynote lectures by prominent researchers and leaders from industry and academia. These keynotes were carefully selected because of their vision, and they met our high expectations. The distinguished keynote speakers list included:

- Rana El-Kaliouby, chief executive officer and cofounder of Affectiva, a startup in the field of emotion intelligence
- Ray Liu, University of Maryland, College Park, and founder of Origin Wireless Inc.
- David Nahamoo, IBM fellow and chief scientist for Conversational Systems
- Jan Rabaey, Donald Oscar Pederson Distinguished Professor, University of California at Berkeley.

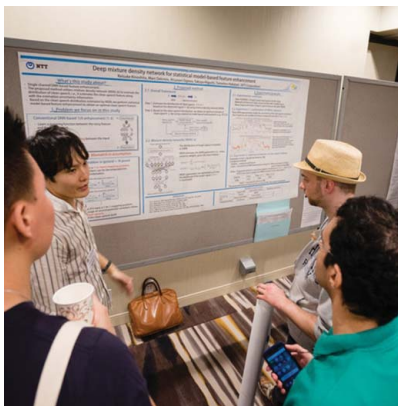
ICASSP 2017 also included several new initiatives. Very popular and warmly welcomed by the attendees, they served as "Tabasco" to add additional seasonings to the ICASSP program. The new initiatives included the following:

- Building on past years' traditions, the student paper contest was held at ICASSP and, for the first time, the contest took place in a dedicated open

Digital Object Identifier 10.1109/MSP.2017.2698119
Date of publication: 11 July 2017

Active Technical Exchanges

Active Technical Exchanges (*continued*)

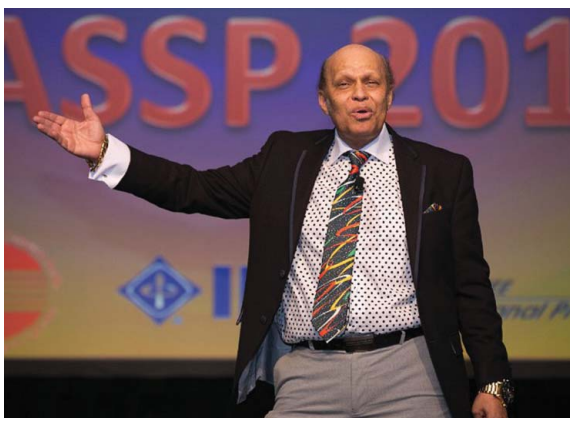


session. The finalist papers were presented at a poster session held on Monday evening, 6 March, during a very relaxed atmosphere of a wine and cheese reception. Top papers received generous awards. In addition, IBM sponsored a student paper award dedicated to a paper in the speech processing area.

- An M.S./Ph.D. forum also took place Monday evening at the same reception. Students had a chance to communicate with professors and industrial professionals. A contest among the presenters took place with awards given to the winners.
- Several events were organized to introduce ICASSP attendees to new and emerging technologies, new research opportunities, and ongoing intellectual discussion:
  - a National Science Foundation panel on funding opportunities in signal processing fields and applications
  - a panel on signal processing open research issues and challenges
  - a forum on acoustics applications and funding opportunities in the Gulf of Mexico for oil spills and environment.
  - the Internet of Things (IoT) in industry forum
  - a workshop on IoT curricula and challenges.

In addition to this stimulating and intellectually inspiring technical program,

Opening Ceremony

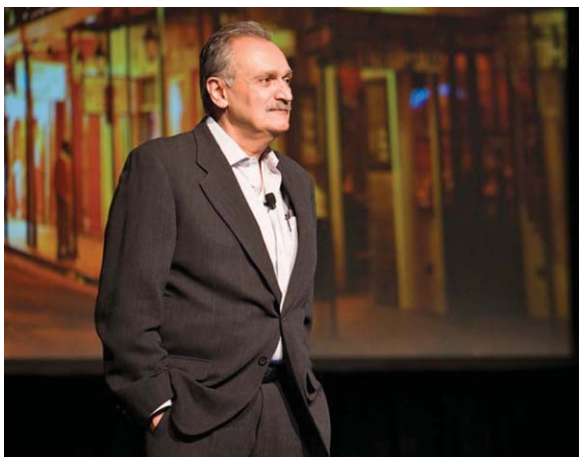


Colorful Social and Cultural Programs

a colorful, cultural, and entertaining social program was planned. Attendees could feel the celebratory mood the entire duration of the conference.

- On Sunday evening, 5 March, a jazz and zydeco festival took place to welcome attendees to ICASSP 2017 and New Orleans. Attendees sampled gumbo, jambalaya, bread pudding, and other New Orleans delicacies. It was also a chance for our community to try the New Orleans two-step and jitterbug. Dance lessons were available to all attendees before the reception.
- On Tuesday evening, 7 March, the Mystick Krewe of ICASSP was

Plenary Speakers



initiated through a Mardi Gras Ball. It was a real Mardi Gras experience topped by the best zydeco band in the world.

- The conference finale took place on 9 March, a goodbye and thanking festival of our own talent in music and singing took place. This talent show fitted very well the spirits of the conference. Students, young professionals, and young spirited friends had a very good time at this social event.

A great time had by all

ICASSP 2017 was a festival of knowledge, information, technology, education, culture, music, and food. We are pleased that 2,231 attendees joined us in New Orleans, and we hope they found ICASSP 2017 a rewarding, informative, and stimulating experience. For readers who could not make this year's conference, we hope you enjoy reading the highlights in this new column in *IEEE Signal Processing Magazine* and will check out the technical results from the ICASSP proceedings in IEEE *Xplore*; the presentation slides and posters archived in IEEE SigPort; and the plenary videos that are available online at http://tiny.cc/ICASSP.

Acknowledgments

Many people, colleagues, and friends helped make this conference a success. Special thanks are given to the SPS leadership, office staff, and administration team. My gratitude is given to all of the members of the organizing committee for their dedication and working as a team as well as Billene Cannon and her conference management service team. I am very grateful to Dr. Savoie, the president of the University of Louisiana-Lafayette and its leadership. I would also like to thank my dear student volunteers who kept the conference running smoothly. Support from the following companies and organizations is gratefully acknowledged: Microsoft Corporation; Alibaba Group; Apple Inc.; Yahoo! Japan Corporation; Baidu Inc.; Google, Inc.; MathWorks; Amazon; Datatang Technology, Inc.; Nuance Communications Inc.; Qualcomm Inc.; IBM Research; and Starkey Hearing Technologies.

SP

DATES AHEAD

Please send calendar submissions to:
Dates Ahead, Att: Jessica Welsh, E-mail: j.welsh@ieee.org

2017

AUGUST

25th European Signal Processing Conference (EUSIPCO)
28 August–2 September, Kos Island, Greece.
General Chairs: Petros Maragos and Sergios Theodoridis
URL: www.eusipco2017.org

14th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)
29 August–1 September, Lecce, Italy.
General Chairs: Cosimo Distante and Larry S. Davis
URL: www.avss2017.org

SEPTEMBER

IEEE International Conference on Image Processing (ICIP)
17–20 September, Beijing, China.
General Chairs: Xinggang Lin, Anthony Vetro, and Min Wu
URL: http://2017.ieeeicip.org/

OCTOBER

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)
15–18 October, New Paltz, New York, USA.
General Chairs: Patrick A. Naylor and Meinard Müller
URL: http://www.waspaa.com/

19th IEEE International Workshop on Multimedia Signal Processing (MMSP)
16–18 October, London-Luton, United Kingdom.
General Chairs: Vladan Velisavljevic, Vladimir Stankovic, and Zixiang Xiong
URL: http://mmsp2017.eee.strath.ac.uk/

Digital Object Identifier 10.1109/MSP.2017.2693144
Date of publication: 11 July 2017

©GRAPHIC STOCK

The main theme of ICME 2017 is "The New Media Experience." Approximately 400 participants mainly from Asia, Europe, and North America will gather in Hong Kong 10–14 July to discuss the latest development in multimedia technologies and related fields.

NOVEMBER

Fifth IEEE Global Conference on Signal and Information Processing (GlobalSIP)
14–16 November 2017, Montréal, Canada.
General Cochairs: Warren Gross and Kostas Plataniotis
URL: http://2017.ieeeglobalsip.org

DECEMBER

Ninth IEEE Workshop on Information Forensics and Security (WIFS)
4–7 December, Rennes, France.
General Cochairs: Teddy Furon and Carmela Troncoso
URL: http://wifs2017.org/

Seventh IEEE Conference of the Sensor Signal Processing for Defence (SSPD)
6–7 December, Edinburgh, Great Britain.
General Chairs: Mike Davies, Jonathon Chambers, and Paul Thomas
URL: www.sspd.eng.ed.ac.uk/

17th IEEE International Workshop on Computational Advances in Multisensor Adaptive Processing (CAMSAP)
10–13 December, Curacao, Dutch Antilles.
General Chairs: André L.F. de Almeida and Martin Haardt
URL: http://www.cs.huji.ac.il/conferences/CAMSAP17/

17th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)
18–20 December, Bilbao, Spain.
General Cochairs: Begona García-Zapirain and Adel Elmaghraby
URL: http://www.isspit.org/isspit/2017/

2018

APRIL

IEEE International Symposium on Biomedical Imaging (ISBI)
4–7 April, Washington, D.C.
Conference Chairs: Amir Amini and Scott Acton
URL: https://biomedicalimaging.org/2018/

43rd International Conference on Acoustic, Speech, and Signal Processing (ICASSP)
22–27 April, Seoul, South Korea.
General Chair: Monson Hayes
General Cochair: Hanseok Ko
URL: http://2018.ieeeicassp.org/

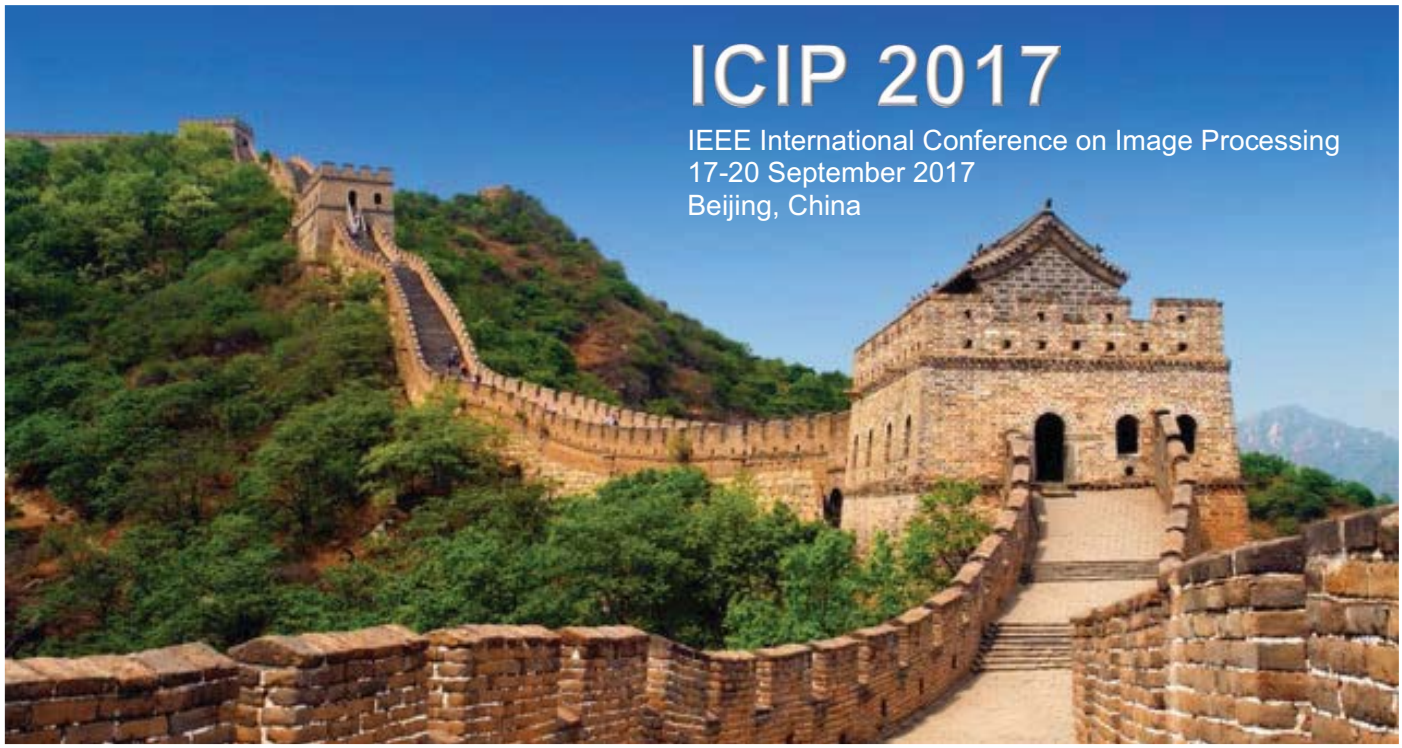
JULY

IEEE International Conference on Multimedia and Expo (ICME)
23–27 July, San Diego, California.
General Chairs: C.-C. Jay Kuo, Truong Nguyen, and Wenjun Zeng
URL: http://www.icme2018.org/

OCTOBER

25th IEEE International Conference on Image Processing (ICIP)
7–10 October, Athens, Greece.

SP



Keynote Speakers

Michael Elad
Prof at Technion

Song-Chun Zhu
Prof at UCLA

Karri Pulli
CTO at Meta

Tutorials

- Vision and Language: Bridging Vision & Language with Deep Learning
- Multi-camera processing, analysis and applications
- Modern First-Order Optimization Methods for Imaging Problems
- Future Video Coding – Coding Tools and Developments beyond HEVC
- Distance Metric Learning for Image and Video Understanding
- Hyperspectral Image and Video Processing
- Scalable Deep Learning for Image Processing

Special Sessions

- Computational Imaging
- Light Field Imaging and Display
- Perceptual Quality Evaluation of Advanced Multimedia Systems
- Saliency Detection and Applications for Image/Video Analysis
- Recent Advances in Video Compression in Open Codecs
- Real-World Visual Content Modeling and Understanding
- Trends in Statistical Analysis of Manifold-Valued Data

Grand Challenges

- Light Field Image Coding
- Video Compression Technology
- Use of Image Restoration for Video Coding
- Content-Based Video Relevance Prediction

Advanced Registration Deadline: July 31
http://2017.ieeeicip.org

Digital Object Identifier 10.1109/MSP.2017.2716389

IN THE SPOTLIGHT

(continued from page 196)

keep some level of anonymity. We need new approaches and tangible solutions to tackle this issue as it will cause significant problems in our future, but the question of how we will accomplish this still remains.

When it comes to secure storage of biometric data, there have been some clever techniques proposed in the past, enabled by signal processing, including fuzzy hash [7] (e.g., the ability to compare two distinctly different items and determine a level of similarity between the two), fuzzy vault [8] (e.g., an encryption scheme that encodes information in a way that is difficult to obtain without a key), and secure sketch [9] techniques. However, these techniques suffer from one of two problems. First, many of the security techniques proposed, from a quantification, storage, and communication point of view, are designed for discrete data and use simple similarity measures. However, true biometric data requires complex similarity functions. Second, the techniques designed for real-world biometric data are either ad hoc and without formal proof of security or don't provide a sufficiently rigorous security formulation.

> **While your face isn't a secret, the data about you and your loved ones that it's linked to should be protected unless we truly do want to live in a "Big Brother" society.**

Is technology giving companies unprecedented access to our data?

For most of us, the use of fingerprints today might be limited to our phone or computer, but what does the future hold for biometric authentication? As technology advances, we will encounter privacy and security issues even more frequently. It's within reach for companies to use new technology to replace all passwords, security personal identification numbers, access codes, etc. MasterCard and HSBC are great examples [10] of companies using facial recognition technology to verify a user's identity. Even Ford is partnering [11] with a machine vision company to add facial recognition technology to its vehicles.

But these advances might allow companies to "go too far" with a person's biometric data, giving unprecedented access. While your face isn't a secret, the data about you and your loved ones that it's linked to should be protected unless we truly do want to live in a "Big Brother" society.

All in all, these security concerns will only increase and evolve with time, but signal processing plays a significant role in providing potential solutions to these issues. Although there is a fascination with the science behind our biometric data, we can't head into a future in which we'll be identified at every step of our lives. We must be diligent in ensuring the right policies and laws prevent biometric data from being used indiscriminately. We must ask ourselves how biometric authentication, which is a convenience in our lives, be prevented from becoming an avenue for companies to invade our privacy.

Author

Nasir Memon (memon@nyu.edu) received his B.E. and M.S. degrees from Birla Institute of Technology and Science, Pilani, India, and his Ph.D. degree from the University of Nebraska. He is a member of the IEEE Signal Processing Society and a professor of computer science and engineering at New York University (NYU) Tandon School of Engineering. He also is an affiliate faculty member in the Computer Science Department in the Courant Institute of Mathematical Sciences at NYU.

References

[1] J. Lynch. (2015). FBI combines civil and criminal fingerprints into one fully searchable database. [Online]. Available: https://www.eff.org/deeplinks/2015/09/little-fanfare-fbi-ramps-biometrics-programs-yet-again

[2] A. Sternstein. (2015). Department of Homeland Security. [Online]. Available: http://www.nextgov.com/defense/2015/01/dhs-launch-iris-and-facial-recognition-border/103908/

[3] Lenovo and Dell—Laptop with fingerprint scanner. [Online]. Available: http://checklaptop.com/best-laptop-with-fingerprint-reader-44/

[4] P. Ausick. (2016). Data breaches top 600 so far in 2016. [Online]. Available: http://247wallst.com/technology-3/2016/08/19/data-breaches-top-600-so-far-in-2016/

[5] M. Snider and E. Weise. (2016). 500 million Yahoo accounts breached. [Online]. Available: http://www.usatoday.com/story/tech/2016/09/22/report-yahoo-may-confirm-massive-data-breach/90824934/

[6] K. Kimachia. (2013). How to protect yourself from unethical or illegal spying. [Online]. Available: http://www.makeuseof.com/tag/how-to-protect-yourself-from-unethical-or-illegal-spying/

[7] (2007). Using fuzzy hashing techniques to identify malicious code. [Online]. Available: http://www.shadowserver.org/wiki/uploads/Information/FuzzyHashing.pdf

[8] (2011). Fuzzy vault. [Online]. Available: https://wiki.cse.buffalo.edu/cse545/content/fuzzy-vault

[9] L. Qiming, S. Yagiz, and M. Nasir, "Secure sketch for biometric templates," *Adv. Cryptology*, vol. 4284, pp. 99–113, 2006. [Online]. Available: http://dx.doi.org/10.1007/11935230_7

[10] T. Wadlow. (2016). HSBC deploys selfie security: Are passwords finished? [Online]. Available: http://www.businessrevieweurope.eu/finance/1048/HSBC-deploys-selfie-security:-are-passwords-finished

[11] J. Carroll. (2016). Ford targeting launch of fully autonomous vehicle by 2021. [Online]. Available: http://www.vision-systems.com/articles/2016/09/ford-targeting-launch-of-fully-autonomous-vehicle-by-2021.html

SP

ADVERTISING & SALES

© GRAPHIC STOCK

The Advertisers' Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine.*

IEEE SIGNAL PROCESSING MAGAZINE REPRESENTATIVE

Mark David, Director, Business Development — Media & Advertising, Phone: +1 732 465 6473, Fax: +1 732 981 1855, m.david@ieee.org

COMPANY	PAGE NUMBER	WEBSITE	PHONE
IEEE USA	5	www.ieeeusa.org/policy/govfel	+1 202 530 8347
Mathworks	CVR 4	www.mathworks.com/wireless	+1 508 647 7040

Digital Object Identifier 10.1109/MSP.2017.2701059

Digital Object Identifier 10.1109/MSP.2017.2715678

Nasir Memon

How Biometric Authentication Poses New Challenges to Our Security and Privacy

The use of biometric data—an individual's measurable physical and behavioral characteristics—isn't new. Government and law enforcement agencies have long used it. The Federal Bureau of Investigation (FBI) has been building a biometric recognition database [1]; the U.S. Department of Homeland Security is sharing [2] its iris and facial recognition of foreigners with the FBI. But the use of biometric data by consumer goods manufacturers for authentication purposes has skyrocketed in recent years. For example, Apple's iPhone allows users to scan their fingerprints to unlock the device, secure mobile bill records, and authenticate payments. Lenovo and Dell [3] leverage fingerprints to enable users to sign onto their computers with just a swipe.

Using biometric data to access our personal devices is increasing as a way to get around the limitations of the commonly used password-based mechanism: it's easier, more convenient, and (theoretically) more secure. But biometric data can also be stolen and used in malicious ways. Capturing fingerprints at scale isn't as easy as lifting a credit card or Social Security number, but experience and history tells us that once something is used extensively, criminals will figure out how to misuse and monetize it.

In addition, with the uptick in data breaches [4] (Yahoo! being the most recent example [5]), we've demonstrated

ADOBE STOCK

we can't keep secrets or properly protect identities. As more companies use biometric authentication, we must be concerned about how our biometric data is secured: currently there is no restriction on what biometric information companies can share and with whom. This is why we need better solutions—we must develop techniques and protocols based on cryptography and signal processing that would protect biometric data and yet allow authentication. We need mechanisms that provide a user some control on when and how their biometric data are being used.

> **As more companies use biometric authentication, we must be concerned about how our biometric data is secured.**

To ensure we're staying on top and ahead of threats to our personal information, we must better understand the dangers associated with the use of biometric authentication (and the role signal processing can play in alleviating them), and the concerns that come to light with technological advances.

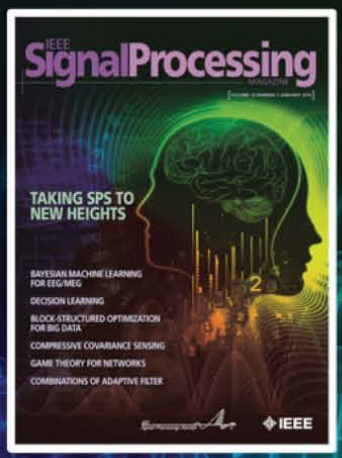
The dangers of frequent biometric authentication

Why is biometric authentication an important issue now more than ever? Companies are increasingly using different means to identify people and assess their buying decisions and how they live their lives. By simply uploading your picture to Facebook or using your thumb to unlock your smartphone, you may be giving away critical data without realizing where the information is going and what it's being used for. It's feasible to envision a society in which we're all identified, all the time and wherever we go. This is dangerous because it can lead to illegal spying [6] from government and law enforcement agencies. To address these concerns, mechanisms must be put in place to permit people to

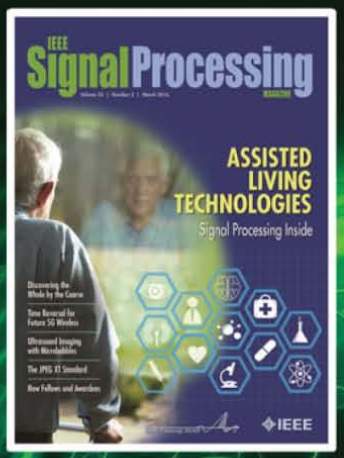
(continued on page 194)

Digital Object Identifier 10.1109/MSP.2017.2697179
Date of publication: 11 July 2017

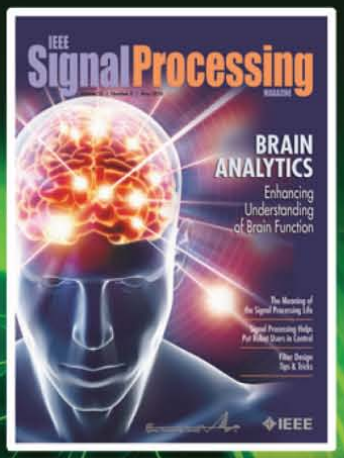
January 2016
Feature Article Collection



March 2016
Assistive Living Technologies



May 2016
Brain Signal Analytics

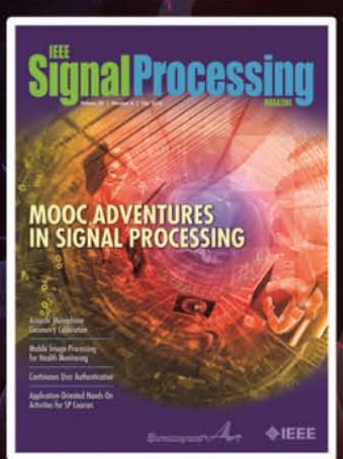


Publish with IEEE Signal Processing Magazine

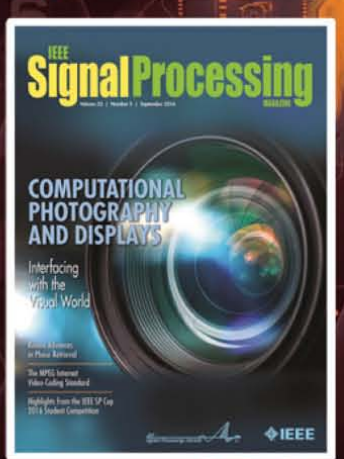
HIGH IMPACT among all electrical engineering publications

REACH a broad signal processing audience worldwide

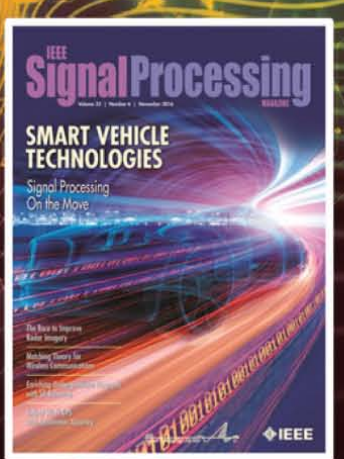
WELCOME proposals for Special Issues and Feature Articles, and contributions to Columns



July 2016
Feature Article Collection



September 2016
Computational Photography & Display



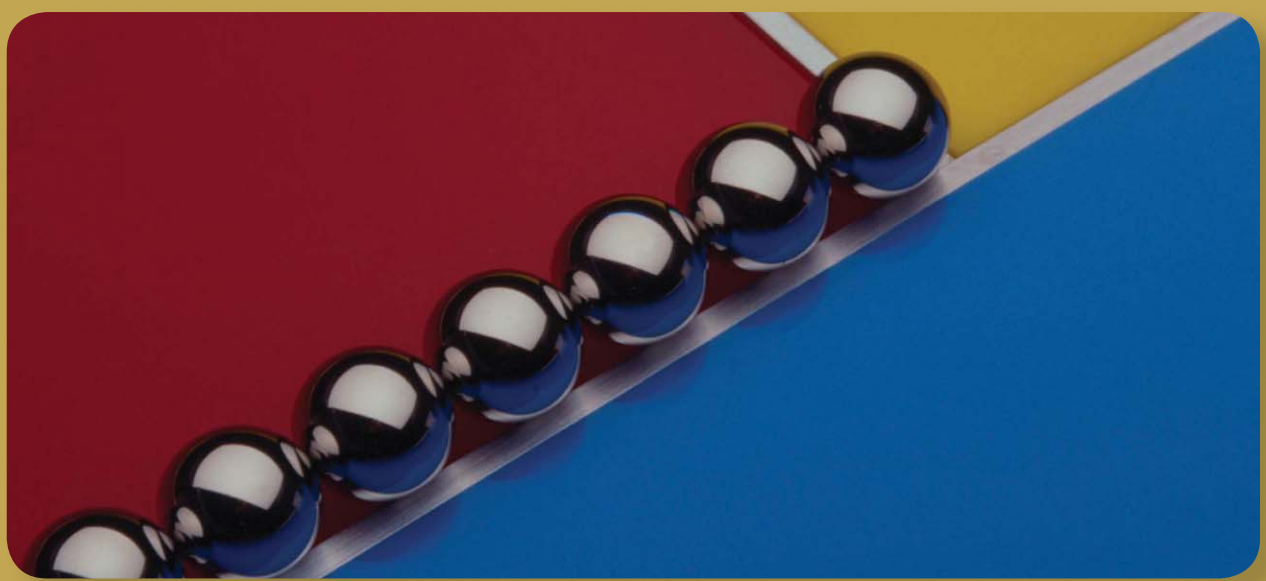
December 2016
SP for Smart Vehicle

Digital Object Identifier 10.1109/MSP.2017.2715698

IEEE SIGNAL PROCESSING SOCIETY

ContentGazette

JULY 2017 ISSN 2167-5023



T-SP July 1 2017 Vol. 65 #13
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7927836

T-SP July 15 2017 Vol. 65 #14
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7935478

T-ASLP June 2017 Vol. 25 #6
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7933040

T-IP June 2017 Vol. 26 #6
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7926471

T-IFS July 2017 Vol. 12 #7
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7951147

T-MM June 2017 Vol.19 #6
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7927828

J-STSP June 2017 Vol. 11 #4
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7927548

T-SPL June 2017 Vol. 24 #6
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7927835

T-CI June 2017 Vol. 3 #2
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7921859

T-SIPN June 2017 Vol. 3 #2
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7931737

http://www.ieee.org/publications_standards/publications/authors/author_ethics.html

IEEE Signal Processing Society

IEEE

Call for Papers and Sponsors

ICASSP2018

The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing

April 22 - 27, 2018, Seoul, Korea

http://2018.ieeeicassp.org

Signal Processing and Artificial Intelligence: Changing the World

Submission of Papers

Authors are invited to submit papers of not more than four pages of technical content including figures and references, with an optional fifth page containing only references. Submission instructions, paper format templates, and other important information will be made available on the ICASSP 2018 website, http://2018.ieeeicassp.org.

Conference Topics

The conference will feature world-class international speakers, tutorials, exhibits, lectures and poster sessions from around the world. Topics include but are not limited to:

- Audio and acoustic signal processing
- Sensor array & multichannel signal processing
- Bio-imaging and biomedical signal processing
- Signal processing education
- Design & implementation of signal processing systems
- Signal processing for communications & networking
- Image, video & multidimensional signal processing
- Signal processing theory & methods
- Industry technology tracks
- Signal processing for big data
- Information forensics and security
- The Internet of Things & RFID
- Machine learning for signal processing
- Speech processing
- Spoken language processing
- Multimedia signal processing
- Remote sensing and signal processing
- Signal processing for brain machine interface
- Signal processing for smart systems
- Signal processing for cyber security
- Computational imaging

Call for Tutorials

Tutorials at ICASSP form an important part of the program, giving attendees the opportunity to learn about current research areas that are of growing interest to the signal processing community. Those who are interested in presenting a tutorial may want to contact one of the tutorial chairs before preparing a formal proposal. It is important to keep in mind, for any tutorial, that it should be tutorial in nature, and within the grasp of a wide audience.

Call for Special Sessions

The program for ICASSP 2018 will include Special Sessions that complement the traditional program with new and emerging topics of significant interest to the signal-processing community, particularly those that are in line with the theme of the conference. Please refer to the conference webpage for information about Special Session proposals.

Call for Exhibitors and Sponsors

ICASSP 2018 offers exhibitors and sponsors an opportunity to showcase their company's products and innovative solutions at the Signal Processing Society's flagship conference that will be held for the first time in the Korean Peninsula. Please refer to the conference webpage for information about signing up to become an exhibitor or sponsor at ICASSP.

Signal Processing Letters

Authors of IEEE Signal Processing Letters (SPL) papers will be given the opportunity to present their work at ICASSP 2018, subject to space availability and approval by the Technical Program Chairs. SPL papers published between January 1, 2017 and December 31, 2017 are eligible for presentation at ICASSP 2018. Because they are already peer-reviewed and published, SPL papers presented at ICASSP 2018 will neither be reviewed nor included in the proceedings.

Important Dates

August 4, 2017
Special Session Proposals Due

August 11, 2017
Tutorial Proposals Due

September 8, 2017
Notification of Special Session Acceptance

September 15, 2017
Notification of Tutorial Acceptance

October 27, 2017
Paper Submissions Due

January 12, 2018
Signal Processing Letters Due

January 26, 2018
Notification of Paper Acceptance

February 9, 2018
Revised Paper Upload Deadline

February 16, 2018
Author Registration Deadline

General Chairs

Monson Hayes
Hanseok Ko

Technical Program Chairs

Dan Schonfeld
Pascale Fung
Nam Ik Cho

Sponsored by

IEEE

IEEE Signal Processing Society

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

JULY 1, 2017 VOLUME 65 NUMBER 13 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

A Unified Successive Pseudoconvex Approximation Framework http://dx.doi.org/10.1109/TSP.2017.2684748 ..... *Y. Yang and M. Pesavento* 3313

Optimal Training Sequences for Large-Scale MIMO-OFDM Systems http://dx.doi.org/10.1109/TSP.2017.2688978 ..... *Z. Sheng, H. D. Tuan, H. H. Nguyen, and M. Debbah* 3329

Cell-Edge-Aware Precoding for Downlink Massive MIMO Cellular Networks http://dx.doi.org/10.1109/TSP.2017.2690387 ..... *H. H. Yang, G. Geraci, T. Q. S. Quek, and J. G. Andrews* 3344

Finite-Resolution Effects in p-Leader Multifractal Analysis http://dx.doi.org/10.1109/TSP.2017.2690391 ..... *R. Leonarduzzi, H. Wendt, P. Abry, S. Jaffard, and C. Melot* 3359

Hourglass Arrays and Other Novel 2-D Sparse Arrays With Reduced Mutual Coupling http://dx.doi.org/10.1109/TSP.2017.2690390 ..... *C.-L. Liu and P. P. Vaidyanathan* 3369

Labeled Random Finite Sets With Moment Approximation http://dx.doi.org/10.1109/TSP.2017.2688960 ..... *Z. Lu, W. Hu, and T. Kirubarajan* 3384

Maximum-Likelihood Approach With Bayesian Refinement for Multichannel-Wiener Postfiltering http://dx.doi.org/10.1109/TSP.2017.2692731 ..... *P. Thüne and G. Enzner* 3399

Discriminative GoDec+ for Classification http://dx.doi.org/10.1109/TSP.2017.2684746 ..... *K. Guo, X. Xu, and D. Tao* 3414

Utility Maximizing Sequential Sensing Over a Finite Horizon http://dx.doi.org/10.1109/TSP.2017.2692725 ..... *L. Ferrari, Q. Zhao, and A. Scaglione* 3430

The Alpha-HMM Estimation Algorithm: Prior Cycle Guides Fast Paths http://dx.doi.org/10.1109/TSP.2017.2692724 ..... *Y. Matsuyama* 3446

Stationary Signal Processing on Graphs http://dx.doi.org/10.1109/TSP.2017.2690388 ..... *N. Perraudin and P. Vandergheynst* 3462

A Scalable Algorithm for Tracking an Unknown Number of Targets Using Multiple Sensors http://dx.doi.org/10.1109/TSP.2017.2688966 ..... *F. Meyer, P. Braca, P. Willett, and F. Hlawatsch* 3478

Fast Approximation Algorithms for a Class of Non-convex QCQP Problems Using First-Order Methods http://dx.doi.org/10.1109/TSP.2017.2690386 ..... *A. Konar and N. D. Sidiropoulos* 3494

Projected Nesterov's Proximal-Gradient Algorithm for Sparse Signal Recovery http://dx.doi.org/10.1109/TSP.2017.2691661 ..... *R. Gu and A. Dogandžić* 3510

Decentralized Clustering and Linking by Networked Agents http://dx.doi.org/10.1109/TSP.2017.2692736 ..... *S. Khawatmi, A. H. Sayed, and A. M. Zoubir* 3526

Robust Control of Varying Weak Hyperspectral Target Detection With Sparse Nonnegative Representation http://dx.doi.org/10.1109/TSP.2017.2688965 ..... *R. Bacher, C. Meillier, F. Chatelain, and O. Michel* 3538

OVERVIEW ARTICLE

Tensor Decomposition for Signal Processing and Machine Learning http://dx.doi.org/10.1109/TSP.2017.2690524 ..... *N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos* 3551

IEEE

(Contents Continued from Page 3309)

Hourglass Arrays and Other Novel 2-D Sparse Arrays With Reduced Mutual Coupling . . . . . *C.-L. Liu and P. P. Vaidyanathan* 3369

Labeled Random Finite Sets With Moment Approximation . . . . . *Z. Lu, W. Hu, and T. Kirubarajan* 3384

Maximum-Likelihood Approach With Bayesian Refinement for Multichannel-Wiener Postfiltering . . . . . *P. Thüne and G. Enzner* 3399

Discriminative GoDec+ for Classification . . . . . *K. Guo, X. Xu, and D. Tao* 3414

Utility Maximizing Sequential Sensing Over a Finite Horizon . . . . . *L. Ferrari, Q. Zhao, and A. Scaglione* 3430

The Alpha-HMM Estimation Algorithm: Prior Cycle Guides Fast Paths . . . . . *Y. Matsuyama* 3446

Stationary Signal Processing on Graphs . . . . . *N. Perraudin and P. Vandergheynst* 3462

A Scalable Algorithm for Tracking an Unknown Number of Targets Using Multiple Sensors . . . . . *F. Meyer, P. Braca, P. Willett, and F. Hlawatsch* 3478

Fast Approximation Algorithms for a Class of Non-convex QCQP Problems Using First-Order Methods . . . . . *A. Konar and N. D. Sidiropoulos* 3494

Projected Nesterov's Proximal-Gradient Algorithm for Sparse Signal Recovery . . . . . *R. Gu and A. Dogandžić* 3510

Decentralized Clustering and Linking by Networked Agents . . . . . *S. Khawatmi, A. H. Sayed, and A. M. Zoubir* 3526

Robust Control of Varying Weak Hyperspectral Target Detection With Sparse Nonnegative Representation . . . . . *R. Bacher, C. Meillier, F. Chatelain, and O. Michel* 3538

OVERVIEW ARTICLE

Tensor Decomposition for Signal Processing and Machine Learning . . . . . *N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos* 3551

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

JULY 15, 2017 VOLUME 65 NUMBER 14 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

Cooperative Simultaneous Localization and Synchronization in Mobile Agent Networks http://dx.doi.org/10.1109/TSP.2017.2691665 ..... *B. Etzlinger, F. Meyer, F. Hlawatsch, A. Springer, and H. Wymeersch* 3587

Adaptive Low-Rank Matrix Completion http://dx.doi.org/10.1109/TSP.2017.2695450 ..... *R. Tripathi, B. Mohan, and K. Rajawat* 3603

Transmit Precoding for Interference Exploitation in the Underlay Cognitive Radio Z-channel http://dx.doi.org/10.1109/TSP.2017.2695448 ..... *K. L. Law, C. Masouros, and M. Pesavento* 3617

Joint Sensing Matrix and Sparsifying Dictionary Optimization for Tensor Compressive Sensing http://dx.doi.org/10.1109/TSP.2017.2699639 ..... *X. Ding, W. Chen, and I. J. Wassell* 3632

An Information Theoretic Approach to Robust Constrained Code Design for MIMO Radars http://dx.doi.org/10.1109/TSP.2017.2692747 ..... *M. M. Naghsh, M. Modarres-Hashemi, M. A. Kerahroodi, and E. H. M. Alian* 3647

Consistent Estimation for Partition-Wise Regression and Classification Models http://dx.doi.org/10.1109/TSP.2017.2698407 ..... *R. C. Y. Cheung, A. Aue, and T. C. M. Lee* 3662

Tensor Decompositions for Identifying Directed Graph Topologies and Tracking Dynamic Networks http://dx.doi.org/10.1109/TSP.2017.2698369 .... *Y. Shen, B. Baingana, and G. B. Giannakis* 3675

Data Sketching for Large-Scale Kalman Filtering http://dx.doi.org/10.1109/TSP.2017.2691662 ..... *D. Berberidis and G. B. Giannakis* 3688

Hankel Matrix Nuclear Norm Regularized Tensor Completion for N-dimensional Exponential Signals http://dx.doi.org/10.1109/TSP.2017.2695566 ... *J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, and X. Qu* 3702

Maximum-Likelihood Detection for MIMO Systems Based on Differential Metrics http://dx.doi.org/10.1109/TSP.2017.2698411 ..... *M.-X. Chang and W.-Y. Chang* 3718

Robust Estimation of Self-Exciting Generalized Linear Models With Application to Neuronal Modeling http://dx.doi.org/10.1109/TSP.2017.2690385 ... *A. Kazemipour, M. Wu, and B. Babadi* 3733

Efficient Techniques for Impulsive Noise Cancellation in CGU/SD Systems http://dx.doi.org/10.1109/TSP.2017.2699645 .... *F. Abdelkefi and J. Ayadi* 3749

An Efficient Global Algorithm for Single-Group Multicast Beamforming http://dx.doi.org/10.1109/TSP.2017.2699640 ..... *C. Lu and Y.-F. Liu* 3761

Low-Computing-Load, High-Parallelism Detection Method Based on Chebyshev Iteration for Massive MIMO Systems With VLSI Architecture http://dx.doi.org/10.1109/TSP.2017.2698410 ..... *G. Peng, L. Liu, P. Zhang, S. Yin, and S. Wei* 3775

Generalized Minimum Noise Subspace For Array Processing http://dx.doi.org/10.1109/TSP.2017.2695457 ..... *V.-D. Nguyen, K. Abed-Meraim, N. Linh-Trung, and R. Weber* 3789

Sparsity-Driven Laplacian-Regularized Outlier Identification for Dictionary Learning http://dx.doi.org/10.1109/TSP.2017.2701310 ..... *P. A. Forero, S. Shafer, and J. D. Harguess* 3803

Exploiting Spatial Channel Covariance for Hybrid Precoding in Massive MIMO Systems http://dx.doi.org/10.1109/TSP.2017.2701321 ..... *S. Park, J. Park, A. Yazdan, and R. W. Heath, Jr.* 3818

Differential Feedback of Geometrical Mean Decomposition Precoder for Time-Correlated MIMO Systems http://dx.doi.org/10.1109/TSP.2017.2692741 ..... *H.-C. Chen and Y.-P. Lin* 3833

Fast Power Allocation for Secure Communication With Full-Duplex Radio http://dx.doi.org/10.1109/TSP.2017.2701318 ..... *L. Chen, Q. Zhu, W. Meng, and Y. Hua* 3846

IEEE

(Contents Continued from Page 3583)

An Information Theoretic Approach to Robust Constrained Code Design for MIMO Radars ..... *M. M. Naghsh, M. Modarres-Hashemi, M. A. Kerahroodi, and E. H. M. Alian* 3647

Consistent Estimation for Partition-Wise Regression and Classification Models ..... *R. C. Y. Cheung, A. Aue, and T. C. M. Lee* 3662

Tensor Decompositions for Identifying Directed Graph Topologies and Tracking Dynamic Networks ..... *Y. Shen, B. Baingana, and G. B. Giannakis* 3675

Data Sketching for Large-Scale Kalman Filtering ..... *D. Berberidis and G. B. Giannakis* 3688

Hankel Matrix Nuclear Norm Regularized Tensor Completion for N-dimensional Exponential Signals ..... *J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, and X. Qu* 3702

Maximum-Likelihood Detection for MIMO Systems Based on Differential Metrics ..... *M.-X. Chang and W.-Y. Chang* 3718

Robust Estimation of Self-Exciting Generalized Linear Models With Application to Neuronal Modeling ..... *A. Kazemipour, M. Wu, and B. Babadi* 3733

Efficient Techniques for Impulsive Noise Cancellation in CGU/SD Systems ..... *F. Abdelkefi and J. Ayadi* 3749

An Efficient Global Algorithm for Single-Group Multicast Beamforming ..... *C. Lu and Y.-F. Liu* 3761

Low-Computing-Load, High-Parallelism Detection Method Based on Chebyshev Iteration for Massive MIMO Systems With VLSI Architecture ..... *G. Peng, L. Liu, P. Zhang, S. Yin, and S. Wei* 3775

Generalized Minimum Noise Subspace For Array Processing ..... *V.-D. Nguyen, K. Abed-Meraim, N. Linh-Trung, and R. Weber* 3789

Sparsity-Driven Laplacian-Regularized Outlier Identification for Dictionary Learning ..... *P. A. Forero, S. Shafer, and J. D. Harguess* 3803

Exploiting Spatial Channel Covariance for Hybrid Precoding in Massive MIMO Systems ..... *S. Park, J. Park, A. Yazdan, and R. W. Heath, Jr.* 3818

Differential Feedback of Geometrical Mean Decomposition Precoder for Time-Correlated MIMO Systems ..... *H.-C. Chen and Y.-P. Lin* 3833

Fast Power Allocation for Secure Communication With Full-Duplex Radio ..... *L. Chen, Q. Zhu, W. Meng, and Y. Hua* 3846

IEEE International Symposium on Signal Processing and Information Technology

December 18-20, 2017 - Bilbao - Spain

ORGANIZERS

General Co-chairs
Begoña García-Zapirain
University of Deusto
Bilbao, Spain

Adel Elmaghraby
University of Louisville
Kentucky, United States

Technical Program Chairs
Ibon Oleagordia Ruiz
Amaia Mendez Zorrila

University of Deusto
Bilbao, Spain

Finance and Registration Chair
Reda Ammar
University of Connecticut
United States

Publication Chair
Daniel Sierra
University of Louisville
Kentucky, USA

Local Arrangements Team
Begoña García Zapirain
Amaia Mendez Zorrilla
Ibon Ruiz Oleagordia
Alvaro Muro
Fernando Hernandez
Alain Sanchez
Christian Castillo
Iranzu Mugueta (Social Events)

ISSPIT 2017 is the 17th IEEE International Symposium on Signal Processing and Information Technology. It is a premiere technical forum for researchers in the fields of signal processing and information technology. ISSPIT 2017 will include state-of-the-art oral, poster sessions, and on-line presentations related to the key areas outlined below. Accepted papers will be published in the Proceedings of IEEE ISSPIT 2017. Best Papers will be considered for Awards and invited to extend the manuscript for journal publication.

The following topics are suggested and other areas may be also considered:

- Signal Processing Theory and Methods
- Signal Processing for Communications and Networking
- Design & Implementation of Signal Processing Systems
- Image, Video & Multidimensional Signal Processing
- Multimedia Signal Processing
- Biological Image and signal processing
- Audio and Acoustic signal Processing
- Health Informatics
- e-Health and m-Health
- Sensor Arrays
- Radar Signal Processing
- Internet Software Architectures
- Multimedia and Image Based Systems
- Mobile Computing and Applications
- e-Commerce
- Bioinformatics and Bioengineering
- Information Processing
- Geographical Information Systems
- Object Based Software Engineering
- Speech Processing
- Computer Networks
- Neural Networks

IMPORTANT DATES
Submission Deadline – September 4, 2017
Acceptance Date – October 15, 2017

Details are on http://www.isspit.org

***Pending IEEE Approval (IEEE SP & IEEE-CS)**

Hosted by:

Deusto
Universidad de Deusto
University of Deusto

DeustoTech

IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society | acm Association for Computing Machinery | www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed | MEDLINE U.S. National Library of Medicine

JUNE 2017 | VOLUME 25 | NUMBER 6 | ITASFA | (ISSN 2329-9290)

SPECIAL SECTION ON SOUND SCENE AND EVENT ANALYSIS

EDITORIAL

Introduction to the Special Section on Sound Scene and Event Analysis http://dx.doi.org/10.1109/TASLP.2017.2699334 ..... *G. Richard, T. Virtanen, J. P. Bello, N. Ono, and H. Glotin* 1169

PAPERS

Maximum Likelihood Decision Fusion for Weapon Classification in Wireless Acoustic Sensor Networks http://dx.doi.org/10.1109/TASLP.2017.2690579 ..... *H. A. Sánchez-Hevia, D. Ayllón, R. Gil-Pita, and M. Rosa-Zurera* 1172

Spectrogram Enhancement Using Multiple Window Savitzky-Golay (MWSG) Filter for Robust Bird Sound Detection http://dx.doi.org/10.1109/TASLP.2017.2690562 ..... *N. R. Koluguri, G. N. Meenakshi, and P. K. Ghosh* 1183
On-Bird Sound Recordings: Automatic Acoustic Recognition of Activities and Contexts http://dx.doi.org/10.1109/TASLP.2017.2690565 .... *D. Stowell, E. Benetos, and L. F. Gill* 1193
Outlier Learning via Augmented Frozen Dictionaries http://dx.doi.org/10.1109/TASLP.2017.2690567 .... *B. T. Carroll, B. M. Whitaker, W. Dayley, and D. V. Anderson* 1207
Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification http://dx.doi.org/10.1109/TASLP.2017.2690570 .... *V. Bisot, R. Serizel, S. Essid, and G. Richard* 1216
Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging http://dx.doi.org/10.1109/TASLP.2017.2690563 .. .. *Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley* 1230
Bag-of-Features Methods for Acoustic Event Detection and Classification http://dx.doi.org/10.1109/TASLP.2017.2690574 .... *R. Grzeszick, A. Plinge, and G. A. Fink* 1242
Supervised Representation Learning for Audio Scene Classification http://dx.doi.org/10.1109/TASLP.2017.2690561 ..... *A. Rakotomamonjy* 1253
Polyphonic Sound Event Tracking Using Linear Dynamical Systems http://dx.doi.org/10.1109/TASLP.2017.2690576 .... *E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley* 1266
Improved Audio Scene Classification Based on Label-Tree Embeddings and Convolutional Neural Networks http://dx.doi.org/10.1109/TASLP.2017.2690564 ..... *H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins* 1278
Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection http://dx.doi.org/10.1109/TASLP.2017.2690575 .... *E. Çakr, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen* 1291
Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016 http://dx.doi.org/10.1109/TASLP.2017.2690569 ..... *J. Schröder, N. Moritz, J. Anemüller, S. Goetze, and B. Kollmeier* 1304
Combining Temporal Features by Local Binary Pattern for Acoustic Scene Classification http://dx.doi.org/10.1109/TASLP.2017.2690558 ... . *W. Yang and S. Krishnan* 1315
Multimodal Kernel Method for Activity Detection of Sound Sources http://dx.doi.org/10.1109/TASLP.2017.2690568 .... *D. Dov, R. Talmon, and I. Cohen* 1322
Spatial Cepstrum as a Spatial Feature Using a Distributed Microphone Array for Acoustic Scene Analysis http://dx.doi.org/10.1109/TASLP.2017.2690559 ..... *K. Imoto and N. Ono* 1335
Robust Detection of Environmental Sounds in Binaural Auditory Scenes http://dx.doi.org/10.1109/TASLP.2017.2690573 .... *I. Trowitzsch, J. Mohr, Y. Kashef, and K. Obermayer* 1344

REGULAR PAPERS

Acoustic Sensor Array Processing

Swarm Intelligence Based Particle Filter for Alternating Talker Localization and Tracking Using Microphone Arrays http://dx.doi.org/10.1109/TASLP.2017.2693566 ..... *K. Wu, V. Gopalan Reju, A. W. H. Khong, and S. T. Goh* 1384

Speaker Recognition and Characterization

DNN-Driven Mixture of PLDA for Robust Speaker Verification http://dx.doi.org/10.1109/TASLP.2017.2692304 .... *N. Li, M.-W. Mak, and J.-T. Chien* 1371

Speech Analysis

Noise Robust Formant Frequency Estimation Method Based on Spectral Model of Repeated Autocorrelation of Speech http://dx.doi.org/10.1109/TASLP.2016.2625423 ..... *A. S. M. M. Jameel, S. A. Fattah, R. Goswami, W.-P. Zhu, and M. O. Ahmad* 1357

EDICS—Editor's Information Classification Scheme http://dx.doi.org/10.1109/TASLP.2017.2705592 1398
Information for Authors http://dx.doi.org/10.1109/TASLP.2017.2705596 1400

IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

JUNE 2017 VOLUME 26 NUMBER 6 IIPRE4 (ISSN 1057-7149)

PAPERS

Fast Unsupervised Bayesian Image Segmentation With Adaptive Spatial Regularisation ..... *M. Pereyra and S. McLaughlin* 2577
Robust Depth-Based Person Re-Identification ..... *A. Wu, W.-S. Zheng, and J.-H. Lai* 2588
Semi-Supervised Multi-View Discrete Hashing for Fast Image Search ..... *C. Zhang and W.-S. Zheng* 2604
Disjunctive Normal Parametric Level Set With Application to Image Segmentation ..... *F. Mesadi, M. Cetin, and T. Tasdizen* 2618
Ultrasound Image Despeckling Using Stochastic Distance-Based BM3D ..... *C. A. N. Santos, D. L. N. Martins, and N. D. A. Mascarenhas* 2632
Rate-Distortion Optimized Graph-Based Representation for Multiview Images With Complex Camera Configurations ..... *X. Su, T. Maugey, and C. Guillemot* 2644
Curl-Constrained Gradient Estimation for Image Recovery From Highly Incomplete Spectral Data ..... *C. Ravazzi, G. Coluccia, and E. Magli* 2656
Multilinear Spatial Discriminant Analysis for Dimensionality Reduction ..... *S. Yuan, X. Mao, and L. Chen* 2669
Enhanced Just Noticeable Difference Model for Images With Pattern Complexity ..... *J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and C.-C. J. Kuo* 2682
A Precision Analysis of Camera Distortion Models .... *Z. Tang, R. Grompone von Gioi, P. Monasse, and J.-M. Morel* 2694
Blind Facial Image Quality Enhancement Using Non-Rigid Semantic Patches ..... *E. Hait and G. Gilboa* 2705
Multi-Scale Multi-Feature Context Modeling for Scene Recognition in the Semantic Manifold ..... *X. Song, S. Jiang, and L. Herranz* 2721

(Contents Continued on Page 2572)

IEEE TRANSACTIONS ON IMAGE PROCESSING (ISSN 1057-7149) is published monthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $631.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in print, electronic, and CD-ROM. **Postmaster**: Send address changes to IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

SignalProcessing | Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page | Qmags

(Contents Continued from Page 2571)

Robust Object Tracking With Discrete Graph-Based Multiple Experts ..... J. Li, C. Deng, R. Y. D. Xu, D. Tao, and B. Zhao 2736
The Recognition of the Point Symbols in the Scanned Topographic Maps ..... Q. Miao, P. Xu, X. Li, J. Song, W. Li, and Y. Yang 2751
GRASS: A Gradient-Based Random Sampling Scheme for Milano Retinex ..... M. Lecca, A. Rizzi, and R. P. Serapioni 2767
Analysis of Packet-Loss-Induced Distortion in View Synthesis Prediction-Based 3D Video Coding ..... P. Gao, Q. Peng, and W. Xiang 2781
Feature Sensitive Label Fusion With Random Walker for Atlas-Based Image Segmentation ..... S. Bao and A. C. S. Chung 2797
Blind Forensics of Successive Geometric Transformations in Digital Images Using Spectral Method: Theory and Applications ..... C. Chen, J. Ni, Z. Shen, and Y. Q. Shi 2811
Deep Label Distribution Learning With Label Ambiguity ..... B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng 2825
Graph Regularized Auto-Encoders for Image Representation ..... Y. Liao, Y. Wang, and Y. Liu 2839
Scale Invariant and Noise Robust Interest Points With Shearlets ..... M. A. Duval-Poo, N. Noceti, F. Odone, and E. De Vito 2853
Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval ..... X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou 2868
Energy-Efficient Images ..... H. Hadizadeh 2882
Classification via Sparse Representation of Steerable Wavelet Frames on Grassmann Manifold: Application to Target Recognition in SAR Image ..... G. Dong, G. Kuang, N. Wang, and W. Wang 2892
Low-Rank Embedding for Robust Image Feature Extraction ..... W. K. Wong, Z. Lai, J. Wen, X. Fang, and Y. Lu 2905
Feature Selection Based on High Dimensional Model Representation for Hyperspectral Images ..... G. Taşkin, H. Kaya, and L. Bruzzone 2918
Dynamical Textures Modeling via Joint Video Dictionary Learning ..... X. Wei, Y. Li, H. Shen, F. Chen, M. Kleinsteuber, and Z. Wang 2929
Clearing the Skies: A Deep Network Architecture for Single-Image Rain Removal ..... X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley 2944
No-Reference Quality Assessment of Tone-Mapped HDR Pictures ..... D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans 2957
Progressive Dictionary Learning With Hierarchical Predictive Structure for Low Bit-Rate Scalable Video Coding ..... W. Dai, Y. Shen, H. Xiong, X. Jiang, J. Zou, and D. Taubman 2972
Structured Sparse Subspace Clustering: A Joint Affinity Learning and Subspace Clustering Framework ..... C.-G. Li, C. You, and R. Vidal 2988
Sequential Dictionary Learning From Correlated Data: Application to fMRI Data Analysis ..... A.-K. Seghouane and A. Iqbal 3002
Re-Weighted Discriminatively Embedded K-Means for Multi-View Clustering ..... J. Xu, J. Han, F. Nie, and X. Li 3016
Deeply Learned View-Invariant Features for Cross-View Action Recognition ..... Y. Kong, Z. Ding, J. Li, and Y. Fu 3028
Optimization of Camera Arrangement Using Correspondence Field to Improve Depth Estimation ..... S. Fu, F. Safaei, and W. Li 3038
Joint Defogging and Demosaicking ..... Y. Lee, K. Hirakawa, and T. Q. Nguyen 3051
Establishing Keypoint Matches on Multimodal Images With Bootstrap Strategy and Global Information ..... Y. Li, H. Jin, J. Wu, and J. Liu 3064

EDICS—Editor's Information Classification Scheme 3077
Information for Authors 3078

IEEE Statistical Signal Processing Workshop 2018
10 – 13 June 2018, Freiburg, Germany
www.ssp2018.org

The 2018 IEEE Workshop on Statistical Signal Processing (SSP) will be held from 10-13 June 2018 in Freiburg, Germany. The SSP Workshop is a unique meeting that brings members of the IEEE Signal Processing Society together with researchers from allied fields such as bioinformatics, communications, machine learning, and statistics. One of its key features is having all contributed and special sessions as poster sessions allowing extensive interaction and networking. The scientific program of SSP 2018 will include invited plenary talks, and regular and special sessions with contributed research papers. All submitted papers are reviewed by experts, and all accepted papers will be published on IEEEXplore.

We invite submitting original research papers on topics including, but not limited to, the following areas:

Foundations, methods, and algorithms
- Detection and estimation theory
- Machine learning and pattern recognition
- Signal separation methods
- Data driven methods
- Bayesian techniques
- Sampling and reconstruction
- Signal and system modeling
- Adaptive signal processing
- Distributed signal processing
- Signal processing over graphs and networks
- Optimization
- Sparsity-aware processing
- Matrix and tensor methods

Application areas
- Bioinformatics and genomics
- Big data
- Signal processing for the internet of things
- Array processing, radar, and sonar
- Communication systems and networks
- Sensor networks
- Information forensics and security
- Medical and biomedical imaging
- Social networks
- Smart grids and industrial applications
- Geoscience
- Astrophysics
- Financial signal processing

Submission of papers: Prospective authors are invited to submit full papers, with up to four pages of technical content (references may be listed on a fifth page), using the template and formatting guidelines posted at www.ssp2018.org. All accepted papers must be presented at the workshop in order to be included in the proceedings. There will be best student paper awards.

Submission of proposals for special sessions: Special session proposals must include a title, rationale, session outline, list of invited papers, and contact information. Please refer to www.ssp2018.org for further information regarding the submission of proposals.

Important dates:

• Submission of proposals for special sessions	24 November 2017
• Notification of acceptance of special sessions	8 December 2017
• Submission of full papers	26 January 2018
• Notification of paper acceptance	23 March 2018
• Registration and camera-ready papers	20 April 2018

Venue: The conference will be held at the spectacular Historical Merchants' Hall, one of the most outstanding historical buildings in Freiburg dating back to the 14th century. It is situated in the historical center of the city right next to Freiburg cathedral and its main square, which features al fresco dining and boutique shopping.

Freiburg is a famous old university town, known for its high standard of living, its beautiful natural setting, and for being the sunniest and warmest city in Germany. It is located in the heart of the Baden wine-growing region close to the Swiss and French borders and serves as the main entry point to the breathtaking beauty of the Black Forest. Freiburg has a convenient high-speed train connection to Frankfurt International Airport.

General Chair
Peter Schreier
Univ. Paderborn

Technical Co-Chairs
Javier Vía
Univ. Cantabria

Arie Yeredor
Tel-Aviv Univ.

Special Sessions
Wing-Kin (Ken) Ma
Chinese Univ. HK

Alle-Jan van der Veen
TU Delft

Finance
Raviv Raich
Oregon State Univ.

Florian Römer
TU Ilmenau

Publications
David Ramírez
Univ. Carlos III Madrid

Local Arrangements
Arno Blau
Stryker Corp

Christian Debes
AGT International

Webmaster
Tim Marrinan
Univ. Paderborn

Publicity/ International Liaison
Tülay Adali
Univ. Maryland BC

Abdelhak Zoubir
TU Darmstadt

NEW!

IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING

30% FASTER TURN AROUND THAN TIP!

Editor-in-Chief

W. Clem Karl
Boston University

Technical Committee

Charles Bouman
Eric Miller
Peter Corcoran
Jong Chul Ye
Dave Brady
William Freeman

The IEEE Transactions on Computational Imaging publishes research results where computation plays an integral role in the image formation process. All areas of computational imaging are appropriate, ranging from the principles and theory of computational imaging, to modeling paradigms for computational imaging, to image formation methods, to the latest innovative computational imaging system designs. Topics of interest include, but are not limited to the following:

Computational Imaging Methods and Models

- Coded image sensing
- Compressed sensing
- Sparse and low-rank models
- Learning-based models, dictionary methods
- Graphical image models
- Perceptual models

Computational Image Formation

- Sparsity-based reconstruction
- Statistically-based inversion methods
- Multi-image and sensor fusion
- Optimization-based methods; proximal iterative methods, ADMM

Computational Photography

- Non-classical image capture
- Generalized illumination
- Time-of-flight imaging
- High dynamic range imaging
- Plenoptic imaging

Computational Consumer Imaging

- Mobile imaging, cell phone imaging
- Camera-array systems
- Depth cameras, multi-focus imaging
- Pervasive imaging, camera networks

Computational Acoustic Imaging

- Multi-static ultrasound imaging
- Photo-acoustic imaging
- Acoustic tomography

Computational Microscopy

- Holographic microscopy
- Quantitative phase imaging
- Multi-illumination microscopy
- Lensless microscopy
- Light field microscopy

Imaging Hardware and Software

- Embedded computing systems
- Big data computational imaging
- Integrated hardware/digital design

Tomographic Imaging

- X-ray CT
- PET
- SPECT

Magnetic Resonance Imaging

- Diffusion tensor imaging
- Fast acquisition

Radar Imaging

- Synthetic aperture imaging
- Inverse synthetic aperture imaging

Geophysical Imaging

- Multi-spectral imaging
- Ground penetrating radar
- Seismic tomography

Multi-spectral Imaging

- Multi-spectral imaging
- Hyper-spectral imaging
- Spectroscopic imaging

For more information on the **IEEE Transactions on Computational Imaging** see *http://www.signalprocessingsociety.org/publications/periodicals/tci/*

IEEE Signal Processing Society ® EMB GRSS

NOW WITH NO PAGE CHARGES!

CALL FOR PAPERS

IEEE Journal of Selected Topics in Signal Processing

Special Issue on Hybrid Analog - Digital Signal Processing for Hardware-Efficient Large Scale Antenna Arrays

5G and beyond systems necessitate the exploitation of high-gain MIMO beamforming/precoding by using large antenna arrays at both the base stations and the mobile units to deliver the high data rates promised. The high cost and power consumption of radio frequency (RF) components such as high-resolution analog-to-digital converters (ADCs) makes dedicating a separate RF chain for each antenna prohibitive, and thus the conventional, fully digital baseband (BB) processing becomes infeasible. This is further pronounced in emerging applications such as the internet of things (IoT) involving massive connectivity. Hybrid analog-digital (AD) processing provides a key solution for allowing a reduced number of RF chains and low-specification RF components, where the transceiver processing is divided into the analog and digital domains. This special issue seeks to bring together contributions from researchers and practitioners in the area of signal processing for wireless communications with an emphasis on new methods for hybrid AD signal processing architectures and transmission. We solicit high-quality original research papers on topics including, but not limited to:

- Fundamental limits of communication by hybrid AD architectures;
- Hybrid AD signal processing techniques for large scale MIMO systems;
- Signal processing techniques robust to low-specification RF components and hardware imperfections;
- Reduced RF chain implementations through parasitic arrays and load modulated MIMO;
- Adaptive transmission / reception techniques for parasitic, reflect, phased, load modulated and other hybrid massive antenna array structures
- Channel modelling for hybrid AD large scale antenna systems;
- Studies and optimization of antenna topologies for massive MIMO deployment with hybrid AD transmission;
- Efficient channel state information (CSI) acquisition techniques for hybrid AD transmission;
- Beamspace MIMO transmission;
- Distributed multi-cell hybrid AD transmission;
- Novel applications of hybrid AD signal processing, including security, energy harvesting, IoT among others;
- Hybrid RF antenna arrays for K, V, W and mmWave frequency bands, including wideband designs;

In addition to technical research results, we invite very high quality submissions of a tutorial or overview nature. We also welcome creative papers outside of the areas listed here but related to the overall scope of the special issue. Prospective authors can contact the Guest Editors to ascertain interest on topics that are not listed above.

Prospective authors should visit http://www.signalprocessingsociety.org/publications/periodicals/jstsp/ for information on paper submission. Manuscripts should be submitted using the Scholar One (Manuscript Central) system at http://mc.manuscriptcentral.com/jstsp-ieee. Manuscripts will be peer reviewed according to the standard IEEE process.

Manuscript Submission:	September 1, 2017
First review completed:	November 1, 2017
Revised manuscript due:	January 1, 2018
Second review completed:	February 1, 2018
Final manuscript due:	March 1, 2018
Publication date:	May 2018

Guest Editors

Dr. Christos Masouros, University College London, UK, email: c.masouros@ucl.ac.uk
Dr. Mathini Sellathurai, Heriot-Watt University, UK, email: m.sellathurai@hw.ac.uk
Prof. Constantinos Papadias, Athens Information Technology, Greece, email: papadias@ait.edu.gr
Prof. Linglong Dai, Tsinghua University, China, email: daill@tsinghua.edu.cn
Prof. Wei Yu, University of Toronto, Canada, email: weiyu@ece.utoronto.ca
Dr. Theodore Sizer, Nokia Bell Labs, U.S.A., email: theodore.sizer@nokia-bell-labs.com

IEEE TRANSACTIONS ON

INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

JULY 2017 VOLUME 12 NUMBER 7 ITIFA6 (ISSN 1556-6013)

REGULAR PAPERS

Fingerprint Recognition of Young Children ..... *A. K. Jain, S. S. Arora, K. Cao, L. Best-Rowden, and A. Bhatnagar* 1501

Security Assurance for System-on-Chip Designs With Untrusted IPs .... *A. Basak, S. Bhunia, T. Tkacik, and S. Ray* 1515

Auditing Anti-Malware Tools by Evolving Android Malware and Dynamic Loading Technique ..... *Y. Xue, G. Meng, Y. Liu, T. H. Tan, H. Chen, J. Sun, and J. Zhang* 1529

BASIS: A Practical Multi-User Broadcast Authentication Scheme in Wireless Sensor Networks ..... *K.-A. Shim* 1545

Stealthy Control Signal Attacks in Linear Quadratic Gaussian Control Systems: Detectability Reward Tradeoff ..... *R. Zhang and P. Venkitasubramaniam* 1555

Local Threshold Design for Target Localization Using Error Correcting Codes in Wireless Sensor Networks in the Presence of Byzantine Attacks ..... *C.-Y. Wei, P.-N. Chen, Y. S. Han, and P. K. Varshney* 1571

NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media ..... *S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi* 1585

Two-Cloud Secure Database for Numeric-Related SQL Range Queries With Privacy Preserving ..... *K. Xue, S. Li, J. Hong, Y. Xue, N. Yu, and P. Hong* 1596

Modeling and Mitigating Impact of False Data Injection Attacks on Automatic Generation Control ..... *R. Tan, H. H. Nguyen, E. Y. S. Foo, D. K. Y. Yau, Z. Kalbarczyk, R. K. Iyer, and H. B. Gooi* 1609

Revisiting Urban War Nibbling: Mobile Passive Discovery of Classic Bluetooth Devices Using Ubertooth One ..... *M. Chernyshev, C. Valli, and M. Johnstone* 1625

(Contents Continued on Page 1498)

IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (ISSN 1556-6013) is published monthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $350.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

(Contents Continued from Page 1497)

Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition Based on Hierarchical MRFs ..... *H. Proença and J. C. Neves* 1637
Face Recognition Using Sparse Fingerprint Classification Algorithm ..... *T. Larrain, J. S. Bernhard, Jr., D. Mery, and K. W. Bowyer* 1646
Pixel-Decimation-Assisted Steganalysis of Synchronize-Embedding-Changes Steganography ..... *S. Tan, H. Zhang, B. Li, and J. Huang* 1658
Research on the Security of Microsoft's Two-Layer Captcha ..... *H. Gao, M. Tang, Y. Liu, P. Zhang, and X. Liu* 1671
Face Verification via Learned Representation on Feature-Rich Video Frames ..... *G. Goswami, M. Vatsa, and R. Singh* 1686
Firewall Fingerprinting and Denial of Firewalling Attacks ..... *A. X. Liu, A. R. Khakpour, J. W. Hulst, Z. Ge, D. Pei, and J. Wang* 1699
Detecting Silicone Mask-Based Presentation Attack via Deep Dictionary Learning ..... *I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar* 1713
A Zero-Leakage Fuzzy Embedder From the Theoretical Formulation to Real Data ..... *G. E. Hine, E. Maiorana, and P. Campisi* 1724
On Optimal PMU Placement-Based Defense Against Data Integrity Attacks in Smart Grid ..... *Q. Yang, D. An, R. Min, W. Yu, X. Yang, and W. Zhao* 1735

IEEE TRANSACTIONS ON MULTIMEDIA

A PUBLICATION OF
THE IEEE CIRCUITS AND SYSTEMS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY

http://www.signalprocessingsociety.org/tmm/

JUNE 2017 | VOLUME 19 | NUMBER 6 | ITREAE | (ISSN 1520-9210)

PAPERS

Compression and Coding

Interview Motion Compensated Joint Decoding for Compressively Sampled Multiview Video Streams ..... *N. Cen, Z. Guan, and T. Melodia* 1117

Online MoCap Data Coding With Bit Allocation, Rate Control, and Motion-Adaptive Post-Processing ..... *C.-H. Kwak and I. V. Bajić* 1127

Image/Video/Graphics Analysis and Synthesis

Fast Image Dehazing Method Based on Linear Transformation ..... *W. Wang, X. Yuan, X. Wu, and Y. Liu* 1142

Weak-Labeled Active Learning With Conditional Label Dependence for Multilabel Image Classification ..... *J. Wu, S. Zhao, V. S. Sheng, J. Zhang, C. Ye, P. Zhao, and Z. Cui* 1156

Large-Scale Multimedia Systems and Benchmarking

Video eCommerce++: Toward Large Scale Online Video Advertising ..... *Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua* 1170

System Design and Optimization

Live Broadcast With Community Interactions: Bottlenecks and Optimizations ..... *X. Ma, C. Zhang, J. Liu, R. Shea, and D. Fu* 1184

Subjective and Objective Quality Assessment, and User Experience

Bayesian Hierarchical Regression Models for QoE Estimation and Prediction in Audiovisual Communications ..... *S. Tasaka* 1195

Multimedia Search and Retrieval

Deep Video Hashing ..... *V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou* 1209

(Contents Continued on Back Cover)

(Contents Continued from Front Cover)

Cross-Modal Retrieval Using Multiordered Discriminative Structured Subspace Learning . . . . . *L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian* 1220
Deep Coupled Metric Learning for Cross-Modal Matching . . . . . *V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou* 1234
Diversified Visual Attention Networks for Fine-Grained Object Classification . . . . . *B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan* 1245
GIFT: Towards Scalable 3D Shape Retrieval . . . . . *S. Bai, X. Bai, Z. Zhou, Z. Zhang, Q. Tian, and L. J. Latecki* 1257
Matryoshka Peek: Toward Learning Fine-Grained, Robust, Discriminative Features for Product Search . . . . . *Z. Kyaw, S. Qi, K. Gao, H. Zhang, L. Zhang, J. Xiao, X. Wang, and T.-S. Chua* 1272
Automatic Synchronization of Multi-user Photo Galleries . . . . . *E. Sansone, K. Apostolidis, N. Conci, G. Boato, V. Mezaris, and F. G. B. De Natale* 1285

Social and Web Multimedia

Who Are Your "Real" Friends: Analyzing and Distinguishing Between Offline and Online Friendships From Social Multimedia Data . . . . . *D. Lu, J. Sang, Z. Chen, M. Xu, and T. Mei* 1299
Two-Stage Friend Recommendation Based on Network Alignment and Series Expansion of Probabilistic Topic Model . . . . . *S. Huang, J. Zhang, D. Schonfeld, L. Wang, and X.-S. Hua* 1314

Real-Time Communications

Perceptual Pruning: A Context-Aware Transcoder for Immersive Video Conferencing Systems . . . . . *P. Pourashraf and F. Safaei* 1327

Error Resilience and Concealment

Sparse Recovery-Based Error Concealment . . . . . *A. Akbari, M. Trocan, and B. Granado* 1339

Media Cloud Computing and Communication

Distributed Compressive Sensing for Cloud-Based Wireless Image Transmission . . . . . *X. Song, X. Peng, J. Xu, G. Shi, and F. Wu* 1351

Multimedia Content Delivery Networks

CrowdTranscoding: Online Video Transcoding With Massive Viewers . . . . . *Q. He, C. Zhang, and J. Liu* 1365

Big Data Analytics On Multimedia Data and Crowd Sourcing for Multimedia Applications

Crowdsourcing Thousands of Specialized Labels: A Bayesian Active Training Approach . . . . . *M. Servajean, A. Joly, D. Shasha, J. Champ, and E. Pacitti* 1376

CORRESPONDENCE

3D Video Signal Processing

Accelerating Image-Domain-Warping Virtual View Synthesis on GPGPU . . . . . *R. Wang, J. Luo, X. Jiang, Z. Wang, W. Wang, G. Li, and W. Gao* 1392

Information for Authors 1401

ANNOUNCEMENT

Introducing IEEE Collabratec 1403

IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society

www.ieee.org/sp/index.html

JUNE 2017 VOLUME 11 NUMBER 4 IJSTGY (ISSN 1932-4553)

ISSUE ON SPOOFING AND COUNTERMEASURES FOR AUTOMATIC SPEAKER VERIFICATION

EDITORIAL

Introduction to the Issue on Spoofing and Countermeasures for Automatic Speaker Verification http://dx.doi.org/10.1109/JSTSP.2017.2698143 ..... *J. Yamagishi, T. H. Kinnunen, N. Evans, P. de Leon, I. Trancoso* 583

PAPERS

ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge http://dx.doi.org/10.1109/JSTSP.2017.2671435 ..... *Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado* 586

Spectral Features for Synthetic Speech Detection http://dx.doi.org/10.1109/JSTSP.2017.2684705 ..... *D. Paul, M. Pal, and G. Saha* 603

Cochlear Filter and Instantaneous Frequency Based Features for Spoofed Speech Detection http://dx.doi.org/10.1109/JSTSP.2017.2647201 ..... *T. B. Patel and H. A. Patil* 616

Front-End for Antispoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition http://dx.doi.org/10.1109/JSTSP.2017.2647202 ..... *K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li* 630

Significance of Source–Filter Interaction for Classification of Natural *vs.* Spoofed Speech http://dx.doi.org/10.1109/JSTSP.2017.2682788 ..... *T. B. Patel and H. A. Patil* 642

Spoofing Speech Detection Using Modified Relative Phase Information http://dx.doi.org/10.1109/JSTSP.2017.2694139 ..... *L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami* 658

Postprocessing Synthetic Speech With a Complex Cepstrum Vocoder for Spoofing Phase-Based Synthetic Speech Detectors http://dx.doi.org/10.1109/JSTSP.2017.2673807 ..... *C. Demiroglu, O. Buyuk, A. Khodabakhsh, and R. Maia* 669

An Investigation of Deep-Learning Frameworks for Speaker Verification Antispoofing http://dx.doi.org/10.1109/JSTSP.2017.2647199 ..... *C. Zhang, C. Yu, and J. H. L. Hansen* 682

Impact of Score Fusion on Voice Biometrics and Presentation Attack Detection in Cross-Database Evaluations http://dx.doi.org/10.1109/JSTSP.2017.2692389 ..... *P. Korshunov and S. Marcel* 693

CALL FOR PAPERS

IEEE Journal of Selected Topics in Signal Processing: Special Issue on Machine Learning for Cognition in Radio Communications and Radar (ML-CR^2)

While machine learning is achieving ground breaking success in speech recognition, computer vision, natural language processing and business analytics, its impact on radio communications, and on the associated problem area in signal processing, has been less pronounced mainly due to the lack of 'big data' and big applications. However, in the era of the Fifth Generation (5G) cellular systems and Internet-of-Things (IoT), some significant changes are under way. For example, as 5G cellular systems demand huge capacity, massive connectivity, high reliability and low latency, acquiring adequate resources to operate such systems is difficult and novel models and algorithms are needed to help improve spectrum utilization by leveraging large-scale databases, full of context and information. These databases can be sourced from handheld devices, network infrastructure, and the environment, such as typical user trajectories provided by vehicular traffic management systems. In addition, government agencies are now willing to share their spectrum with commercial users. The 3550-3650 MHz band is identified for spectrum sharing between military radars and communication systems. This requires cognition both in communication systems and radars. There is also a general trend toward cognitive radars as the next generation of environment-adaptive radars with unprecedented spectral and behavioral agility. A natural approach to handling all this is the development of efficient machine learning algorithms, which, combined with traditional signal processing methods, will allow for the automation of cognitive functionality both in radars and radio communication networks. There are nontrivial challenges and open questions in the application of machine learning to RF environments starting with the fact that, as opposed to speech recognition and computer vision where the output of machine cognition can be readily compared and verified against human auditory and visual perception, no such option is available for radio signals. The main goal of this Special Issue is to raise awareness of this emerging interdisciplinary research area, and to showcase the existing state-of-the-art and its current and future challenges. Topics of interest include (but are not limited to):

- Machine learning for blind channel and signal characterization
- Machine learning for source separation
- Deep learning for RF signal classification
- Machine learning for channel decoding
- Machine learning for RF-based geolocation and signal association
- Distributed multi-agent learning in collaborative radio networks
- Machine learning-based antenna selection
- Reinforcement Learning in wireless networks
- Machine learning of the topology and structural properties of radio networks
- Joint optimization and learning of spectrum usage dynamics and spectrum access control
- Privacy-preserving machine learning for cognitive radio
- Machine learning for cognitive technologies in 5G cellular networks
- Non-parametric Bayesian machine learning for temporal clustering of spectral activity
- Machine learning for activity recognition of partially observable wireless network nodes
- Machine learning in cognitive radars for spectrum sharing with communication devices
- Machine learning for passive radars
- Machine learning for Bayesian target characterization
- Machine learning for cognitive radar characterization and for radar waveform design

Prospective authors should follow the instructions given on the IEEE JSTSP webpages: https://signalprocessingsociety.org/publications-resources/ieee-journal-selected-topics-signal-processing and submit their manuscript with the web submission system at: https://mc.manuscriptcentral.com/jstsp-ieee.

Dates:

Manuscript submission: ~~June 30, 2017~~ **July 14, 2017**

First review completed: August 15, 2017

Revised Manuscript Due: September 15, 2017

Second Review Completed: October 31, 2017

Final Manuscript Due: December 15, 2017

Publication: February 2018

Guest Editors (ge.ml.crcr@gmail.com):

Maria Sabrina Greco, University of Pisa

Silvija Kokalj-Filipovic, U.S. Naval Research Laboratory

H. Vincent Poor, Princeton University

George Stantchev, U. S. Naval Research Laboratory

Liang Xiao, Xiamen University

CALL FOR PAPERS

IEEE Journal of Selected Topics in Signal Processing

Special Issue on Hybrid Analog - Digital Signal Processing for Hardware-Efficient Large Scale Antenna Arrays

5G and beyond systems necessitate the exploitation of high-gain MIMO beamforming/precoding by using large antenna arrays at both the base stations and the mobile units to deliver the high data rates promised. The high cost and power consumption of radio frequency (RF) components such as high-resolution analog-to-digital converters (ADCs) makes dedicating a separate RF chain for each antenna prohibitive, and thus the conventional, fully digital baseband (BB) processing becomes infeasible. This is further pronounced in emerging applications such as the internet of things (IoT) involving massive connectivity. Hybrid analog-digital (AD) processing provides a key solution for allowing a reduced number of RF chains and low-specification RF components, where the transceiver processing is divided into the analog and digital domains. This special issue seeks to bring together contributions from researchers and practitioners in the area of signal processing for wireless communications with an emphasis on new methods for hybrid AD signal processing architectures and transmission. We solicit high-quality original research papers on topics including, but not limited to:

- Fundamental limits of communication by hybrid AD architectures;
- Hybrid AD signal processing techniques for large scale MIMO systems;
- Signal processing techniques robust to low-specification RF components and hardware imperfections;
- Reduced RF chain implementations through parasitic arrays and load modulated MIMO;
- Adaptive transmission / reception techniques for parasitic, reflect, phased, load modulated and other hybrid massive antenna array structures
- Channel modelling for hybrid AD large scale antenna systems;
- Studies and optimization of antenna topologies for massive MIMO deployment with hybrid AD transmission;
- Efficient channel state information (CSI) acquisition techniques for hybrid AD transmission;
- Beamspace MIMO transmission;
- Distributed multi-cell hybrid AD transmission;
- Novel applications of hybrid AD signal processing, including security, energy harvesting, IoT among others;
- Hybrid RF antenna arrays for K, V, W and mmWave frequency bands, including wideband designs;

In addition to technical research results, we invite very high quality submissions of a tutorial or overview nature. We also welcome creative papers outside of the areas listed here but related to the overall scope of the special issue. Prospective authors can contact the Guest Editors to ascertain interest on topics that are not listed above.

Prospective authors should visit http://www.signalprocessingsociety.org/publications/periodicals/jstsp/ for information on paper submission. Manuscripts should be submitted using the Scholar One (Manuscript Central) system at http://mc.manuscriptcentral.com/jstsp-ieee. Manuscripts will be peer reviewed according to the standard IEEE process.

Manuscript Submission:	September 1, 2017
First review completed:	November 1, 2017
Revised manuscript due:	January 1, 2018
Second review completed:	February 1, 2018
Final manuscript due:	March 1, 2018
Publication date:	May 2018

Guest Editors

Dr. Christos Masouros, University College London, UK, email: c.masouros@ucl.ac.uk
Dr. Mathini Sellathurai, Heriot-Watt University, UK, email: m.sellathurai@hw.ac.uk
Prof. Constantinos Papadias, Athens Information Technology, Greece, email: papadias@ait.edu.gr
Prof. Linglong Dai, Tsinghua University, China, email: daill@tsinghua.edu.cn
Prof. Wei Yu, University of Toronto, Canada, email: weiyu@ece.utoronto.ca
Dr. Theodore Sizer, Nokia Bell Labs, U.S.A., email: theodore.sizer@nokia-bell-labs.com

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

JUNE 2017 VOLUME 24 NUMBER 6 ISPLEM (ISSN 1070-9908)

LETTERS

SkeletonNet: Mining Deep Part Features for 3-D Action Recognition ..... *Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid* 731
Deep Learning for Quality Assessment in Live Video Streaming ..... *M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta* 736
Explicit APM-LDPC Codes With Girths 6, 8, and 10 ..... *M. Gholami and M. Alinia* 741
Cramér–Rao Bounds for a Coupled Mixture of Polynomial Phase and Sinusoidal FM Signals ..... *P. Wang, P. V. Orlik, K. Sadamoto, W. Tsujita, and Y. Sawa* 746
Postcapture Focusing Using Regression Forest ..... *G. Kanojia and S. Raman* 751
Contravariant Adaptation on the Manifold of Invertible Matrix Transfer Functions ..... *T. K. Moon and J. H. Gunther* 756
No-Reference JPEG Image Quality Assessment Based on Blockiness and Luminance Change .... *Y. Zhan and R. Zhang* 760
A Graphical Evolutionary Game Approach to Social Learning ..... *X. Cao and K. J. R. Liu* 765
New Word Extraction From Chinese Financial Documents ..... *L. Yan, B. Bai, W. Chen, and D. O. Wu* 770
Fast Inverse-Free Sparse Bayesian Learning via Relaxed Evidence Lower Bound Maximization ..... *H. Duan, L. Yang, J. Fang, and H. Li* 774
Joint Transmit and Receive Beamforming for Hybrid Active–Passive Radar ..... *Y. Gao, H. Li, and B. Himed* 779
Iterative Soft/Hard Thresholding With Homotopy Continuation for Sparse Recovery ..... *Y. Jiao, B. Jin, and X. Lu* 784
A Scaling-Less Newton–Raphson Pipelined Implementation for a Fixed-Point Reciprocal Operator ..... *E. Libessart, M. Arzel, C. Lahuec, and F. Andriulli* 789
Efficient Continuous Beam Steering for Planar Arrays of Differential Microphones ..... *A. Bernardini, M. D'Aria, R. Sannino, and A. Sarti* 794
Consistent Change Point Detection for Piecewise Constant Signals With Normalized Fused LASSO ..... *A. Owrang, M. Malek-Mohammadi, A. Proutiere, and M. Jansson* 799

(Contents Continued on Page 728)

IEEE SIGNAL PROCESSING LETTERS (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. NJ Telephone: +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $357.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Postmaster:** Send address changes to IEEE SIGNAL PROCESSING LETTERS, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

(Contents Continued from Page 727)

Contrast Enhancement Using Combined 1-D and 2-D Histogram-Based Techniques . . . . . *D. Kim and C. Kim* 804
Suppressing Random Artifacts in Reference Sensor Pattern Noise via Decorrelation . . . . . *Q. Rao and J. Wang* 809
Unitary Algorithm for Nonseparable Linear Canonical Transforms Applied to Iterative Phase Retrieval . . . . . *L. Zhao, J. T. Sheridan, and J. J. Healy* 814
A Direct-Path Interference Resistant Passive Detector . . . . . *X. Zhang, H. Li, and B. Himed* 818
A Coherence-Based Algorithm for Optimizing Rank-1 Grassmannian Codebooks . . . . *H. E. A. Laue and W. P. du Plessis* 823
Local Adaptive Binary Patterns Using Diamond Sampling Structure for Texture Classification . . . . . *Z. Pan, X. Wu, Z. Li, and Z. Zhou* 828
Iterative Decision Feedback Equalization for SC-FDE Systems Without Fine Timing Synchronization . . . . . *K. Zhang, H. Yu, Y. Hu, and Z. Shen* 833
Plug-In Measure-Transformed Quasi-Likelihood Ratio Test for Random Signal Detection . . . . . *N. Halay and K. Todros* 838
A New Contourlet Transform With Adaptive Directional Partitioning . . . . . *H. Zhao, X. Zhao, T. Zhang, and Y. Liu* 843
Light-Field Image Super-Resolution Using Convolutional Neural Network . . . . . *Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon* 848
Adaptive Metric Learning and Probe-Specific Reranking for Person Reidentification . . . . . *Y. Xie, H. Yu, X. Gong, and M. D. Levine* 853
On the Convergence of Constrained Particle Filters . . . . . *N. Amor, N. C. Bouaynaya, R. Shterenberg, and S. Chebbi* 858
Efficient Compressed Sensing for Wireless Neural Recording: A Deep Learning Approach . . . . . *B. Sun and H. Feng* 863
Subjective Assessment of Super Multiview Video with Coding Artifacts . . . . . *R. Recio, P. Carballeira, J. Gutiérrez, and N. García* 868
What to Expect When You Are Expecting on the Grassmannian . . . . . *A. Eftekhari, L. Balzano, and M. B. Wakin* 872
Wavelet-Based Total Variation and Nonlocal Similarity Model for Image Denoising . . . . . *Y. Shen, Q. Liu, S. Lou, and Y.-L. Hou* 877
Efficient Stepsize Selection Strategy for Givens Parametrized ICA Applied to EEG Denoising . . . . . *M. Saleh, A. Karfoul, A. Kachenoura, I. Merlet, and L. Albera* 882
Asymptotic and Bootstrap Tests for the Dimension of the Non-Gaussian Subspace . . . . . *K. Nordhausen, H. Oja, D. E. Tyler, and J. Virta* 887
Toward High-Quality Real-Time Signal Reconstruction From STFT Magnitude . . . . . *Z. Prša and P. Rajmic* 892
Analysis of the FFT-FBMC Equalization in Selective Channels . . . . . *R. Zakaria and D. Le Ruyet* 897
An Efficient Weighted Least Squares Estimator for Elliptic Localization in Distributed MIMO Radars . . . . . *R. Amiri and F. Behnia* 902
Discriminative Bag-of-Words-Based Adaptive Appearance Model for Robust Visual Tracking . . . . . *F. Zeng, Z. Huang, and Y. Ji* 907
Novel Fractional-Order Difference Schemes Reducible to Standard Integer-Order Formulas . . . . . *M. P. Paskaš, I. S. Reljin, and B. D. Reljin* 912
Proactive Eavesdropping in Relaying Systems . . . . . *X. Jiang, H. Lin, C. Zhong, X. Chen, and Z. Zhang* 917

EDICS—Editor's Information Classification Scheme 922
Information for Authors 923

IEEE TRANSACTIONS ON

COMPUTATIONAL IMAGING

A PUBLICATION OF
IEEE SIGNAL PROCESSING SOCIETY
IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY
IEEE CONSUMER ELECTRONICS SOCIETY

TECHNICALLY CO-SPONSORED BY
IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY

JUNE 2017 VOLUME 3 NUMBER 2 ITCIAJ (ISSN 2333-9403)

GUEST EDITORIAL

Computational Imaging for Earth Sciences ..... *S. Aeron, E. L. Miller, M. Crawford, A. Malcom, A. Reigber, and J. Chanussot* 144

SPECIAL ISSUE ON COMPUTATIONAL IMAGING FOR EARTH SCIENCES

Fast Hyperspectral Unmixing in Presence of Nonlinearity or Mismodeling Effects ..... *A. Halimi, J. M. Bioucas-Dias, N. Dobigeon, G. S. Buller, and S. McLaughlin* 146

Distributed Blind Hyperspectral Unmixing via Joint Sparsity and Low-Rank Constrained Non-Negative Matrix Factorization ..... *C. G. Tsinos, A. A. Rontogiannis, and K. Berberidis* 160

Robust Fusion of Multiband Images With Different Spatial and Spectral Resolutions for Change Detection ..... *V. Ferraris, N. Dobigeon, Q. Wei, and M. Chabert* 175

Fast and Accurate Multiplicative Decomposition for Fringe Removal in Interferometric Images ..... *D.-C. Soncco, C. Barbanson, M. Nikolova, A. Almansa, and Y. Ferrec* 187

Colored Coded Aperture Design in Compressive Spectral Imaging via Minimum Coherence ..... *A. Parada-Mayorga and G. R. Arce* 202

Sensing Matrix Design via Mutual Coherence Minimization for Electromagnetic Compressive Imaging Applications ..... *R. Obermeier and J. A. Martinez-Lorenzo* 217

(Contents Continued on Next Page)

(Contents Continued from Previous Page)

Large-Scale Feature Selection With Gaussian Mixture Models for the Classification of High Dimensional Remote Sensing Images . . . . . *A. Lagrange, M. Fauvel, and M. Grizonnet* 230
Unsupervised Data Driven Feature Extraction by Means of Mutual Information Maximization . . . . . *A. Marinoni and P. Gamba* 243
Online Target Recognition for Time-Sensitive Space Information Networks . . . . . *C. Huo, Z. Zhou, K. Ding, and C. Pan* 254
Beating Level-Set Methods for 5-D Seismic Data Interpolation: A Primal-Dual Alternating Approach . . . . . *R. Kumar, O. López, D. Davis, A. Y. Aravkin, and F. J. Herrmann* 264
Low-Rank Decomposition Based on Disjoint Component Analysis With Applications in Seismic Imaging . . . . . *K. Nose-Filho and J. M. T. Romano* 275
Convex Recovery From Interferometric Measurements . . . . . *L. Demanet and V. Jugnon* 282
Non-Linear Inverse Scattering via Sparsity Regularized Contrast Source Inversion . . . . . *M. T. Bevacqua, L. Crocco, L. D. Donato, and T. Isernia* 296
Salt Reconstruction in Full-Waveform Inversion With a Parametric Level-Set Method . . . . . *A. Kadu, T. van Leeuwen, and W. A. Mulder* 305
Fast GPU-Based Seismogram Simulation From Microseismic Events in Marine Environments Using Heterogeneous Velocity Models . . . . . *S. Das, X. Chen, and M. P. Hobson* 316
On Acoustic Signal Compression for Ultrasonic Borehole Imaging . . . . . *B. Fan, S. Aeron, A. Pedrycz, and H.-P. Valero* 330
The Effect of Hardware-Computed Travel Time on Localization Accuracy in the Inversion of Experimental (Acoustic) Waveform Data . . . . . *M. Takala, T. D. Hämäläinen, and S. Pursiainen* 344
Sparse Clustered Bayesian-Inspired $T_1 - T_2$ Inversion From Borehole NMR Measurements . . . . . *P. Wang, L. Venkataramanan, and V. Jain* 355
High-Fidelity Real-Time Imaging With Electromagnetic Logging-While-Drilling Measurements . . . . . *M. Thiel and D. Omeragic* 369

EDICS—Editor's Classification Information Scheme 379

Information for Authors 380

IEEE TRANSACTIONS ON
SIGNAL AND INFORMATION PROCESSING OVER NETWORKS

A PUBLICATION OF
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY

IEEE Signal Processing Society | IEEE COMMUNICATIONS SOCIETY | IEEE computer society

JUNE 2017 | VOLUME 3 | NUMBER 2 | ITSIBW | (ISSN 2373-776X)

SPECIAL SECTION ON DISTRIBUTED INFORMATION PROCESSING IN SOCIAL NETWORKS

EDITORIAL

Introduction to the Issue on Distributed Information Processing in Social Networks http://dx.doi.org/10.1109/TSIPN.2017.2699868 ..... *Z. Zhang, W. P. Tay, M. Draief, X. Wang, E. K. P. Chong, and A. O. Hero III* 219
Social Learning Over Weakly Connected Graphs http://dx.doi.org/10.1109/TSIPN.2017.2668138 ..... *H. Salami, B. Ying, and A. H. Sayed* 222
Information Cascades With Noise http://dx.doi.org/10.1109/TSIPN.2017.2682798 ..... *T. N. Le, V. G. Subramanian, and R. A. Berry* 239
Analyzing Resilience of Interest-Based Social Networks Against Node and Link Failures http://dx.doi.org/10.1109/TSIPN.2017.2697827 ... *J. Zhao* 252
Data Trading With Multiple Owners, Collectors, and Users: An Iterative Auction Mechanism http://dx.doi.org/10.1109/TSIPN.2017.2668144 ..... *X. Cao, Y. Chen, and K. J. R. Liu* 268
Preserving Privacy Enables "Coexistence Equilibrium" of Competitive Diffusion in Social Networks http://dx.doi.org/10.1109/TSIPN.2017.2697819 ..... *J. Zhao and J. Zhang* 282
Tracking Infection Diffusion in Social Networks: Filtering Algorithms and Threshold Bounds http://dx.doi.org/10.1109/TSIPN.2017.2697940 ..... *V. Krishnamurthy, S. Bhatt, and T. Pedersen* 298
Temporally Agnostic Rumor-Source Detection http://dx.doi.org/10.1109/TSIPN.2017.2668141 ..... *A. Kumar, V. S. Borkar, and N. Karamchandani* 316
Hidden Chinese Restaurant Game: Grand Information Extraction for Stochastic Network Learning http://dx.doi.org/10.1109/TSIPN.2017.2682799 ..... *C.-Y. Wang, Y. Chen, and K. J. R. Liu* 330
Detecting Changes in Dynamic Events Over Networks http://dx.doi.org/10.1109/TSIPN.2017.2696264 ..... *S. Li, Y. Xie, M. Farajtabar, A. Verma, and L. Song* 346
Stochastic Multidimensional Scaling http://dx.doi.org/10.1109/TSIPN.2017.2668145 ..... *K. Rajawat and S. Kumar* 360
Posting Behavior Dynamics and Active Filtering for Content Diversity in Social Networks http://dx.doi.org/10.1109/TSIPN.2017.2696738 ..... *A. R. Masson, Y. Hayel, and E. Altman* 376
Multireceiver Predicate Encryption for Online Social Networks http://dx.doi.org/10.1109/TSIPN.2017.2697580 ..... *C.-I. Fan, Y.-F. Tseng, J.-J. Huang, S.-F. Chen, and H. Kikuchi* 388

IEEE International Symposium on Biomedical Imaging

April 4-7, 2018, Omni Shoreham Hotel, Washington, D.C.

The ***IEEE International Symposium on Biomedical Imaging (ISBI)*** is a scientific conference dedicated to mathematical, algorithmic, and computational aspects of biological and biomedical imaging, across all scales of observation. It fosters knowledge transfer among different imaging communities and contributes to an integrative approach to biomedical imaging. ISBI is a joint initiative from the IEEE Signal Processing Society (SPS) and the IEEE Engineering in Medicine and Biology Society (EMBS). The 2018 meeting will include tutorials, and a scientific program composed of plenary talks, invited special sessions, challenges, as well as oral and poster presentations of peer-reviewed papers. High-quality papers are requested containing original contributions to the topics of interest including image formation and reconstruction, computational and statistical image processing and analysis, dynamic imaging, visualization, image quality assessment, and physical, biological, and statistical modeling. Accepted 4-page regular papers will be published in the symposium proceedings published by IEEE and included in IEEE Xplore. To encourage attendance by a broader audience of imaging scientists and offer additional presentation opportunities, ISBI 2018 will continue to have a second track featuring posters selected from 1-page abstract submissions without subsequent archival publication.

Keynote Speakers

Sam Achilefu, Washington University in St. Louis, Optical Molecular Imaging
Kim Butts Pauly, Stanford University, MR guided Focused Ultrasound
Peter Basser, National Institutes of Health, MR Diffusion Tensor Imaging
Anne Carpenter, Broad Institute, Harvard/MIT, Cellular Image Analysis
Laura Waller, UC Berkeley, Microscopy and Computational Imaging

Important Dates

Proposal Submission	
Tutorials, Special Sessions, & Challenges	*May 29, 2017- September 25, 2017*
4 page papers	
Submission Opens	*July 31, 2017*
Submission Deadline	*October 16, 2017*
Accept/Reject Notification	*December 18, 2017*
Final Submission	*January 8, 2018*
1 page papers	
Submission Opens	*November 20, 2017*
Submission Deadline	*December 18, 2017*
Accept/Reject Notification	*January 22, 2018*
Final Submission Opens	*January 22, 2018*
Final Submission Deadline	*February 5, 2018*

Organizing Committee:

Conference Chairs
Amir Amini *(amir.amini@louisville.edu)*
Scott Acton *(acton@virginia.edu)*

Program Chairs
Erik Meijering *(meijering@imagescience.org)*
Ron Summers *(rms@nih.gov)*

Plenary Chairs
Julia Schnabel *(julia.schnabel@kcl.ac.uk)*
Michael Unser *(michael.unser@epfl.ch)*

Tutorial Chairs
Jerry Prince *(prince@jhu.edu)*
Tanveer Syeda-Mahmood *(stf@us.ibm.com)*

Special Sessions
Arrate Muñoz-Barrutia *(mamunozb@ing.uc3m.es)*
Erin Girard *(erin.girard@siemens.com)*

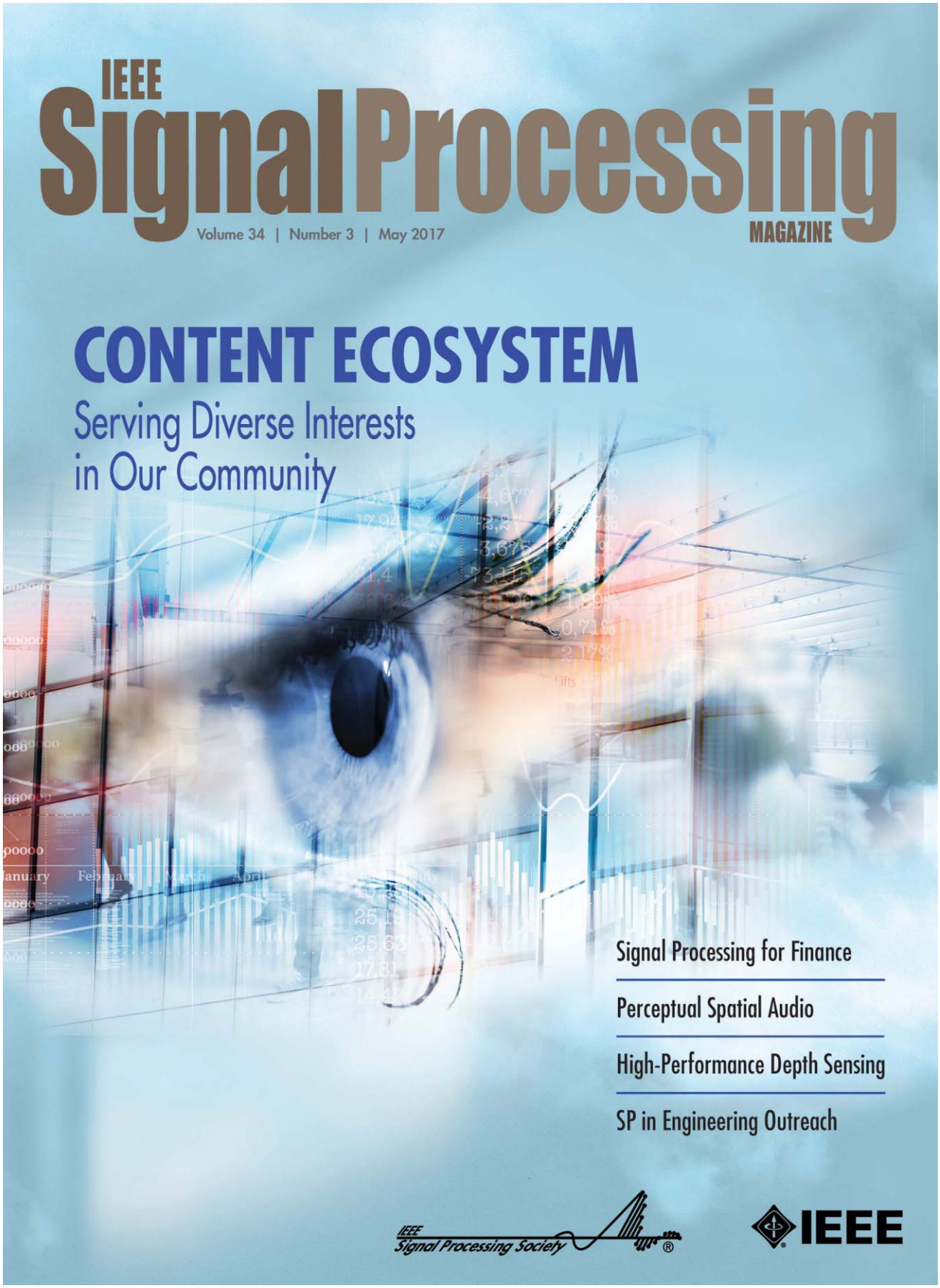
Challenges
Bram van Ginneken *(bram.vanginneken@radboudumc.nl)*
Stephen Aylward *(stephen.aylward@kitware.com)*
Dan Weller *(dsw8c@virginia.edu)*

Exhibits and Industry Liasons
Marie-Pierre Jolly *(marie-pierre.jolly@siemens. com)*
Erhan Bas *(base@janelia.hhmi.org)*
Eliot Siegel *(esiegel@umaryland.edu)*

Student Awards
Yuping Wang *(wyp@tulane.edu)*
Lei Ying *(leiying@buffalo.edu)*

Local Arrangements
Siddhartha Sikdar *(ssikdar@gmu.edu)*
Juan Cebral *(jcebral@gmu.edu)*
Murray Loew *(loew@gwu.edu)*

http://biomedicalimaging.org/2018/


IEEE
Signal Processing
MAGAZINE
Volume 34 | Number 3 | May 2017
CONTENT ECOSYSTEM
Serving Diverse Interests in Our Community
Signal Processing for Finance
Perceptual Spatial Audio
High-Performance Depth Sensing
SP in Engineering Outreach
IEEE Signal Processing Society
IEEE

Contents

Volume 34 | Number 3 | May 2017

FEATURES

14 **SIGNAL PROCESSING FOR FINANCE, ECONOMICS, AND MARKETING**
Xiao-Ping (Steven) Zhang and Fang Wang

36 **PERCEPTUAL SPATIAL AUDIO RECORDING, SIMULATION, AND RENDERING**
Hüseyin Hacıhabiboğlu, Enzo De Sena, Zoran Cvetković, James Johnston, and Julius O. Smith III

55 **COMPUTATIONAL DEPTH SENSING**
Zhiwei Xiong, Yueyi Zhang, Feng Wu, and Wenjun Zeng

ON THE COVER

This issue presents feature articles showcasing the significant and versatile roles that signal processing has been playing in quantitative finance, spatial audio, and visual depth sensing. The "SP Education" column presents systematic engineering outreach efforts at precollegiate levels, for which signal processing is a key building block in many projects engaging young students. The "Life Sciences" column presents velocity estimation in medical ultrasound images. The "From the Editor" column highlights an effort of "content ecosystem" that the magazine is exploring with other Society initiatives.

COVER IMAGE: ©ISTOCKPHOTO.COM/RZOZE19

81 **Lecture Notes**
The CNN as a Guided Multilayer RECOS Transform
C.-C. Jay Kuo

Effectively Interpreting Discrete Wavelet Transformed Signals
Rodrigo Capobianco Guido

94 **Life Sciences**
Velocity Estimation in Medical Ultrasound
Jørgen Arendt Jensen, Carlos Armando Villagómez Hoyos, Simon Holbek, and Kristoffer Lindskov Hansen

101 **Tips & Tricks**
Autocalibrated Sampling Rate Conversion in the Frequency Domain
Lifan Zhao, Xiumei Li, Lu Wang, and Guoan Bi

112 **In the Spotlight**
Signal Processing for Social Good
Kush R. Varshney


PG. 112

COLUMNS

6 **Reader's Choice**
Top Downloads in IEEE *Xplore*

8 **Society News**
New Society Officers Elected for 2018

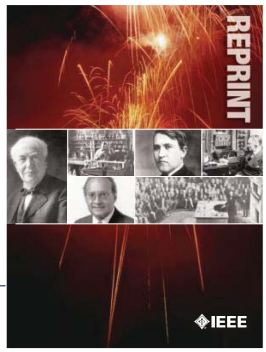
10 **Special Reports**
New Directions in Navigation and Positioning
John Edwards

69 **SP Education**
Engineering Outreach: Yesterday, Today, and Tomorrow
Mónica F. Bugallo and Angela M. Kelly


PG. 69

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US$20.00 (first copy only), nonmembers US$241.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. **For all other copying, reprint, or republication permission,** write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 **Printed in the U.S.A.**

Digital Object Identifier 10.1109/MSP.2017.2678378



IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print is easy, cost-effective and quick.
Complete this form, send via our secure fax (24 hours a day) to 732-981-8062 or mail it back to us.

PLEASE FILL OUT THE FOLLOWING

Author: _____

Publication Title: _____

Paper Title: _____

_____

RETURN THIS FORM TO:
IEEE Publishing Services
445 Hoes Lane
Piscataway, NJ 08855-1331

Email the Reprint Department at reprints@ieee.org for questions regarding this form

PLEASE SEND ME

■ 50 ■ 100 ■ 200 ■ 300 ■ 400 ■ 500 or _____(in multiples of 50) reprints.
■ YES ■ NO Self-covering/title page required. COVER PRICE: $74 per 100, $39 per 50.
■ $58.00 Air Freight must be added for all orders being shipped outside the U.S.
■ $21.50 must be added for all USA shipments to cover the cost of UPS shipping and handling.

PAYMENT

■ Check enclosed. Payable on a bank in the USA.
■ Charge my: ■ Visa ■ Mastercard ■ Amex ■ Diners Club

Account # _____ Exp. date _____

Cardholder's Name (please print): _____

_____

■ Bill me (you must attach a purchase order) Purchase Order Number _____

Send Reprints to:

_____
_____
_____
_____

Bill to address, if different:

_____
_____
_____
_____

Because information and papers are gathered from various sources, there may be a delay in receiving your reprint request. This is especially true with postconference publications. Please provide us with contact information if you would like notification of a delay of more than 12 weeks.

Telephone: _____ Fax: _____ Email Address: _____

2012 REPRINT PRICES (without covers)

Number of Text Pages

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
50	$129	$213	$245	$248	$288	$340	$371	$408	$440	$477	$510	$543
100	$245	$425	$479	$495	$573	$680	$742	$817	$885	$953	$1021	$1088

Larger quantities can be ordered. Email reprints@ieee.org with specific details.

Tax Applies on shipments of regular reprints to CA, DC, FL, MI, NJ, NY, OH and Canada (GST Registration no. 12534188).
Prices are based on black & white printing. Please call us for full color price quote, if applicable.

Authorized Signature: _____ Date: _____

IEEE
Advancing Technology for Humanity

2017 IEEE MEMBERSHIP APPLICATION

(students and graduate students must apply online)

Start your membership immediately: Join online www.ieee.org/join

Please complete both sides of this form, typing or **printing in capital letters**. Use only English characters and abbreviate only if more than 40 characters and spaces per line. We regret that incomplete applications cannot be processed.

1 Name & Contact Information

Please PRINT your name as you want it to appear on your membership card and IEEE correspondence. As a key identifier for the IEEE database, circle your last/surname.

☐ Male ☐ Female Date of birth (Day/Month/Year) ____/____/____

Title First/Given Name Middle Last/Family Surname

▼ Primary Address ☐ Home ☐ Business (All IEEE mail sent here)

Street Address

City State/Province

Postal Code Country

Primary Phone

Primary E-mail

▼ Secondary Address ☐ Home ☐ Business

Company Name Department/Division

Street Address City State/Province

Postal Code Country

Secondary Phone

Secondary E-mail

To better serve our members and supplement member dues, your postal mailing address is made available to carefully selected organizations to provide you with information on technical services, continuing education, and conferences. Your e-mail address is not rented by IEEE. Please check box only if you do not want to receive these postal mailings to the selected address. ☐

2 Attestation

I have graduated from a three- to five-year academic program with a university-level degree.
☐ Yes ☐ No

This program is in one of the following fields of study:
☐ Engineering
☐ Computer Sciences and Information Technologies
☐ Physical Sciences
☐ Biological and Medical Sciences
☐ Mathematics
☐ Technical Communications, Education, Management, Law and Policy
☐ Other (please specify): _____

This academic institution or program is accredited in the country where the institution is located. ☐ Yes ☐ No ☐ Do not know

I have _____ years of professional experience in teaching, creating, developing, practicing, or managing within the following field:

☐ Engineering
☐ Computer Sciences and Information Technologies
☐ Physical Sciences
☐ Biological and Medical Sciences
☐ Mathematics
☐ Technical Communications, Education, Management, Law and Policy
☐ Other (please specify): _____

3 Please Tell Us About Yourself

Select the numbered option that best describes yourself. This information is used by IEEE magazines to verify their annual circulation. Please enter numbered selections in the boxes provided.

A. Primary line of business → ☐

1. Computers
2. Computer peripheral equipment
3. Software
4. Office and business machines
5. Test, measurement and instrumentation equipment
6. Communications systems and equipment
7. Navigation and guidance systems and equipment
8. Consumer electronics/appliances
9. Industrial equipment, controls and systems
10. ICs and microprocessors
11. Semiconductors, components, sub-assemblies, materials and supplies
12. Aircraft, missiles, space and ground support equipment
13. Oceanography and support equipment
14. Medical electronic equipment
15. OEM incorporating electronics in their end product (not elsewhere classified)
16. Independent and university research, test and design laboratories and consultants (not connected with a mfg. co.)
17. Government agencies and armed forces
18. Companies using and/or incorporating any electronic products in their manufacturing, processing, research or development activities
19. Telecommunications services, telephone (including cellular)
20. Broadcast services (TV, cable, radio)
21. Transportation services (airline, railroad, etc.)
22. Computer and communications and data processing services
23. Power production, generation, transmission and distribution
24. Other commercial users of electrical, electronic equipment and services (not elsewhere classified)
25. Distributor (reseller, wholesaler, retailer)
26. University, college/other educational institutions, libraries
27. Retired
28. Other_____

B. Principal job function → ☐

1. General and corporate management
2. Engineering management
3. Project engineering management
4. Research and development management
5. Design engineering management—analog
6. Design engineering management—digital
7. Research and development engineering
8. Design/development engineering—analog
9. Design/development engineering—digital
10. Hardware engineering
11. Software design/development
12. Computer science
13. Science/physics/mathematics
14. Engineering (not elsewhere specified)
15. Marketing/sales/purchasing
16. Consulting
17. Education/teaching
18. Retired
19. Other _____

C. Principal responsibility → ☐

1. Engineering and scientific management
2. Management other than engineering
3. Engineering design
4. Engineering
5. Software: science/mngmnt/engineering
6. Education/teaching
7. Consulting
8. Retired
9. Other _____

D. Title → ☐

1. Chairman of the Board/President/CEO
2. Owner/Partner
3. General Manager
4. VP Operations
5. VP Engineering/Dir. Engineering
6. Chief Engineer/Chief Scientist
7. Engineering Management
8. Scientific Management
9. Member of Technical Staff
10. Design Engineering Manager
11. Design Engineer
12. Hardware Engineer
13. Software Engineer
14. Computer Scientist
15. Dean/Professor/Instructor
16. Consultant
17. Retired
18. Other _____

Are you now or were you ever a member of IEEE?
☐ Yes ☐ No If yes, provide, if known:

Membership Number Grade Year Expired

4 Please Sign Your Application

I hereby apply for IEEE membership and agree to be governed by the IEEE Constitution, Bylaws, and Code of Ethics. I understand that IEEE will communicate with me regarding my individual membership and all related benefits. **Application must be signed.**

Signature Date

Over Please

(continued on next page)

5 Add IEEE Society Memberships (Optional)

The 39 IEEE Societies support your technical and professional interests. Many society memberships include a personal subscription to the core journal, magazine, or newsletter of that society. **For a complete list of everything included with your IEEE Society membership, visit www.ieee.org/join.** All prices are quoted in US dollars.

Please check ✔ the appropriate box.

		BETWEEN 16 AUG 2016-28 FEB 2017 PAY	BETWEEN 1 MAR 2017-15 AUG 2017 PAY
IEEE Aerospace and Electronic Systems ⌘ 🖂	AES010	25.00 ☐	12.50 ☐
IEEE Antennas and Propagation ⌘ 🖂	AP003	15.00 ☐	7.50 ☐
IEEE Broadcast Technology ▪ ⌘ 🖂	BT002	15.00 ☐	7.50 ☐
IEEE Circuits and Systems ▪ ⌘ 🖂	CAS004	22.00 ☐	11.00 ☐
IEEE Communications ⌘ 🖂	COM019	33.00 ☐	16.50 ☐
IEEE Components, Packaging, & Manu. Tech. ▪ ⌘ 🖂	CPMT021	15.00 ☐	7.50 ☐
IEEE Computational Intelligence ⌘ 🖂	CIS011	29.00 ☐	14.50 ☐
IEEE Computer ▪ ⌘ 🖂	C016	60.00 ☐	30.00 ☐
IEEE Consumer Electronics ▪ ⌘ 🖂	CE008	20.00 ☐	10.00 ☐
IEEE Control Systems ⌘ 🖂	CS023	25.00 ☐	12.50 ☐
IEEE Dielectrics and Electrical Insulation ⌘ 🖂	DEI032	26.00 ☐	13.00 ☐
IEEE Education ⌘ 🖂	E025	20.00 ☐	10.00 ☐
IEEE Electromagnetic Compatibility ▪ ⌘ 🖂	EMC027	31.00 ☐	15.50 ☐
IEEE Electron Devices ▪ ⌘ 🖂	ED015	18.00 ☐	9.00 ☐
IEEE Engineering in Medicine and Biology ⌘ 🖂	EMB018	40.00 ☐	20.00 ☐
IEEE Geoscience and Remote Sensing ⌘ 🖂	GRS029	19.00 ☐	9.50 ☐
IEEE Industrial Electronics ⌘ 🖂	IE013	9.00 ☐	4.50 ☐
IEEE Industry Applications ⌘ 🖂	IA034	20.00 ☐	10.00 ☐
IEEE Information Theory ▪ ⌘ 🖂	IT012	30.00 ☐	15.00 ☐
IEEE Instrumentation and Measurement ⌘ 🖂	IM009	29.00 ☐	14.50 ☐
IEEE Intelligent Transportation Systems ▪ ⌘ 🖂	ITSS038	35.00 ☐	17.50 ☐
IEEE Magnetics ▪ 🖂 ♦⌘	MAG033	26.00 ☐	13.00 ☐
IEEE Microwave Theory and Techniques ⌘ 🖂	MTT017	24.00 ☐	12.00 ☐
IEEE Nuclear and Plasma Sciences ▪ ⌘ 🖂	NPS005	35.00 ☐	17.50 ☐
IEEE Oceanic Engineering ▪ ⌘ 🖂	OE022	19.00 ☐	9.50 ☐
IEEE Photonics ▪ ⌘ 🖂	PHO036	34.00 ☐	17.00 ☐
IEEE Power Electronics ⌘ 🖂	PEL035	25.00 ☐	12.50 ☐
IEEE Power & Energy ▪ ⌘ 🖂	PE031	35.00 ☐	17.50 ☐
IEEE Product Safety Engineering ▪	PSE043	35.00 ☐	17.50 ☐
IEEE Professional Communication ▪ ⌘ 🖂	PC026	31.00 ☐	15.50 ☐
IEEE Reliability ▪ ⌘ 🖂	RL007	35.00 ☐	17.50 ☐
IEEE Robotics and Automation ⌘ 🖂	RA024	9.00 ☐	4.50 ☐
IEEE Signal Processing ▪ ⌘ 🖂	SP001	22.00 ☐	11.00 ☐
IEEE Social Implications of Technology ▪ ⌘ 🖂	SIT030	33.00 ☐	16.50 ☐
IEEE Solid-State Circuits ⌘ 🖂	SSC037	22.00 ☐	11.00 ☐
IEEE Systems, Man, & Cybernetics ⌘ 🖂	SMC028	12.00 ☐	6.00 ☐
IEEE Technology & Engineering Management ⌘ 🖂	TEM014	35.00 ☐	17.50 ☐
IEEE Ultrasonics, Ferroelectrics, & Frequency Control ▪ ⌘ 🖂	UFFC020	20.00 ☐	10.00 ☐
IEEE Vehicular Technology ▪ ⌘ 🖂	VT006	18.00 ☐	9.00 ☐

Legend—Society membership includes:

⌘ One or more Society publications
▪ Society newsletter
🖂 Online access to publication
♦ CD-ROM of selected society publications

Complete both sides of this form, sign, and return to:
IEEE MEMBERSHIP APPLICATION PROCESSING
445 HOES LN, PISCATAWAY, NJ 08854-4141 USA
or fax to +1 732 981 0225
or join online at www.ieee.org/join

Please reprint your full name here

6 2017 IEEE Membership Rates (student rates available online)

IEEE member dues and regional assessments are based on where you live and when you apply. Membership is based on the calendar year from 1 January through 31 December. All prices are quoted in US dollars.

Please check ✔ the appropriate box.

RESIDENCE	BETWEEN 16 AUG 2016-28 FEB 2017 PAY	BETWEEN 1 MAR 2017-15 AUG 2017 PAY
United States	$199.00 ☐	$99.50 ☐
Canada (NB, NF, NS, and PEI HST)*	$190.35 ☐	$95.18 ☐
Canada (ON HST)	$187.37 ☐	$93.69 ☐
Canada (GST)*	$175.45 ☐	$87.73 ☐
Canada (GST and QST Quebec)	$190.31 ☐	$95.16 ☐
Africa, Europe, Middle East	$162.00 ☐	$81.00 ☐
Latin America	$153.00 ☐	$76.50 ☐
Asia, Pacific	$154.00 ☐	$77.00 ☐

*IEEE Canada Business No. 125634188

Minimum Income or Unemployed Provision

Applicants who certify that their prior year income did not exceed US$14,900 (or equivalent) or were not employed are granted 50% reduction in: full-year dues, regional assessment and fees for one IEEE Membership plus one Society Membership. If applicable, please check appropriate box and adjust payment accordingly. Student members are not eligible.

☐ I certify I earned less than US$14,900 in 2016
☐ I certify that I was unemployed in 2016

7 More Recommended Options

Proceedings of the IEEE ..... print $49.00 ☐ **or** online $43.00 ☐
Proceedings of the IEEE (print/online combination) ..... $59.00 ☐
IEEE Standards Association (IEEE-SA) ..... $54.00 ☐
IEEE Women in Engineering (WIE) ..... $25.00 ☐

8 Payment Amount

Please total the Membership dues, Society dues, and other amounts from this page:

IEEE Membership dues ⑥ ..... $_____
IEEE Society dues (optional) ⑤ ..... $_____
IEEE-SA/WIE dues (optional) ⑦ ..... $_____
Proceedings of the IEEE (optional) ⑦ ..... $_____
Canadian residents pay 5% GST or appropriate HST (BC–12%; ON–13%; NB, NF, NS, PEI–15%) on Society payments & publications only ..... TAX $_____

AMOUNT PAID ..... **TOTAL** $_____

Payment Method

All prices are quoted in US dollars. You may pay for IEEE membership by credit card (see below), check, or money order payable to IEEE, drawn on a US bank.

☐ Check ☐ VISA ☐ MasterCard ☐ Discover ☐ American Express ☐ Discover

Credit Card Number

MONTH EXPIRATION DATE | YEAR | CARDHOLDER'S 5-DIGIT ZIP CODE (BILLING STATEMENT ADDRESS) USA ONLY

Name as it appears on card

Signature

Auto Renew my Memberships and Subscriptions (available when paying by credit card).
☐ I agree to the Terms and Conditions located at www.ieee.org/autorenew

9 Were You Referred to IEEE?

☐ Yes ☐ No If yes, provide the following:

Member Recruiter Name _____

IEEE Recruiter's Member Number (Required) _____

CAMPAIGN CODE	PROMO CODE

16-MEM-031 P 6/16

Information for Authors

(Updated/Effective January 2015)

For Transactions and Journals:

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Comment Correspondences (brief items that provide comment on a paper previously published in these TRANSACTIONS).

Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must:

i. provide a clear statement of the problem and what the contribution of the work is to the relevant research community;
ii. state why this contribution is significant (what impact it will have);
iii. provide citation of the published literature most closely related to the manuscript; and
iv. state what is distinctive and new about the current manuscript relative to these previously published works.

By submission of your manuscript to these TRANSACTIONS, all listed authors have agreed to the authorship list and all the contents and confirm that the work is original and that figures, tables and other reported results accurately reflect the experimental work. In addition, the authors all acknowledge that they accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended. The corresponding author is responsible for obtaining consent from all co-authors and, if needed, from sponsors before submission.

In order to be considered for review, a paper must be within the scope of the journal and represent a novel contribution. A paper is a candidate for an Immediate Rejection if it is of limited novelty, e.g. a straightforward combination of theories and algorithms that are well established and are repeated on a known scenario. Experimental contributions will be rejected without review if there is insufficient experimental data. These TRANSACTIONS are published in English. Papers that have a large number of typographical and/or grammatical errors will also be rejected without review.

In addition to presenting a novel contribution, acceptable manuscripts must describe and cite related work in the field to put the contribution in context. Do not give theoretical derivations or algorithm descriptions that are easily found in the literature; merely cite the reference.

New and revised manuscripts should be prepared following the "Manuscript Submission" guidelines below, and submitted to the online manuscript system, ScholarOne Manuscripts. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will access your manuscript electronically via the ScholarOne Manuscript system.

Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts.* If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
2. *Electronic Manuscript.* Prepare a PDF file containing your manuscript in double-column, single-spaced format using a font size of 10 points or larger, having a margin of at least 1 inch on all sides. Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOne-Manuscripts site. Since many reviewers prefer a larger font, you are strongly encouraged to also submit a single-column, double-spaced version (11 point font or larger), which is easy to create with the templates provided **IEEE Author Digital Toolbox (http://www.ieee.org/publications_standards/publications/authors/authors_journals.html)**. Page length restrictions will be determined by the double-column version. Proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT" your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a .PDF or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.
3. *EDICS (Not applicable to Journal of Selected Topics in Signal Processing).* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at the at the publication's EDICS webpage (*please see the list below). Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
4. *Additional Documents for Review.* Please upload pdf versions of all items in the reference list that are not publicly available, such as unpublished (submitted) papers. Graphical abstracts and supplemental materials intended to appear with the final paper (see below) must also be uploaded for review at the time of the initial submission for consideration in the review process. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
5. *Supplemental Materials.* IEEE Xplore can publish multimedia files (audio, images, video), datasets, and software (e.g. Matlab code) along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see IEEE Author Digital Toolbox under "Multimedia." To make your work reproducible by others, these TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper.
6. *Submission.* After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
7. *Copyright Form and Consent Form.* By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE. The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Comment Correspondence. Comment Correspondences provide brief comments on material previously published in these TRANSACTIONS. These items may not exceed 2 pages in double-column, single spaced format, using 9 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Correspondence items are submitted in the same way as regular manuscripts (see "Manuscript Submission" above for instructions).

Authors may also submit manuscripts of overview articles, but note that these include an additional white paper approval process http://www.signalprocessingsociety.org/publications/overview-articles/. *[This does not apply to the Journal of Selected Topics in Signal Processing. Please contact the Editor-in-Chief.]*

Digital Object Identifier

Manuscript Length. For the initial submission of a regular paper, the manuscript may not exceed 13 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. Supplemental materials and graphical abstracts are not included in the page count. For regular papers, the revised manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. For Overview Papers, the maximum length is double that for regular submissions at each stage (please reference http://www.signalprocessingsociety.org/publications/overview-articles/ for more information).

Note that any paper in excess of 10 pages will be subject to mandatory overlength page charges. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions. Note: Papers submitted to the TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges.

Exceptions to manuscript length requirements may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all overlength or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of manuscripts rejected from any journal are allowed to resubmit their manuscripts only once. At the time of submission, you will be asked whether your manuscript is a new submission or a resubmission of an earlier rejected manuscript. If it is a resubmission of a manuscript previously rejected by any journal, you are expected to submit supporting documents identifying the previous submission and detailing how your new version addresses all of the reviewers' comments. Papers that do not disclose connection to a previously rejected paper or that do not provide documentation as to changes made may be immediately rejected.

Author Misconduct. Author misconduct includes plagiarism, self-plagiarism, and research misconduct, including falsification or misrepresentation of results. All forms of misconduct are unacceptable and may result in sanctions and/or other corrective actions. Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an author's previously published work that also involves other authors. Self-plagiarism involves the verbatim copying or reuse of an authors own prior work without appropriate citation, including duplicate submission of a single journal manuscript to two different journals, and submission of two different journal manuscripts which overlap substantially in language or technical contribution. For more information on the definitions, investigation process, and corrective actions related to author misconduct, see the Signal Processing Society Policies and Procedures Manual, Section 6.1. http://www.signalprocessingsociety.org/about-sps/governance/policy-procedure/part-2. Author misconduct may also be actionable by the IEEE under the rules of Member Conduct.

Extensions of the Author's Prior Work. It is acceptable for conference papers to be used as the basis for a more fully developed journal submission. Still, authors are required to cite their related prior work; the papers cannot be identical; and the journal publication must include substantively novel aspects such as new experimental results and analysis or added theoretical work. The journal publication should clearly specify how the journal paper offers novel contributions when citing the prior work. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper, and the prior work must be cited as the primary source.

Submission Format. Authors are required to prepare manuscripts employing the on-line style files developed by IEEE, which include guidelines for abbreviations, mathematics, and graphics. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at the **IEEE Author Digital Toolbox** under "Template for all TRANSACTIONS." (LaTeX and MS Word). Please note the following requirements about the abstract:

- The abstract must be a concise yet comprehensive reflection of what is in your article.
- The abstract must be self-contained, without abbreviations, footnotes, displayed equations, or references.
- The abstract must be between 150-250 words.
- The abstract should include a few keywords or phrases, as this will help readers to find it. Avoid over-repetition of such phrases as this can result in a page being rejected by search engines.

In addition to written abstracts, papers may include a graphical abstract; see http://www.ieee.org/publications_standards/publications/authors/authors_journals.html. for options and format requirements.

IEEE supports the publication of author names in the native language alongside the English versions of the names in the author list of an article. For more information, see "Author names in native languages" (http://www.ieee.org/publications_standards/publications/authors/auth_names_native_lang.pdf) on the IEEE Author Digital Toolbox page.

Open Access. The publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted $1,750 OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. Any other application charges (such as overlength page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of $110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (two pages, in the case of Correspondences).

Mandatory Page Charges The author(s) or his/her/their company or institution will be billed $220 per each page in excess of the first ten published pages for regular papers and six published pages for correspondence items. (**NOTE: Papers accepted to IEEE TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges.) These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt. No mandatory overlength page charges will be applied to overview articles in the Society's journals.

Color Charges. Color figures which appear in color only in the electronic (Xplore) version can be used free of charge. In this case, the figure will be printed in the hardcopy version in grayscale, and the author is responsible that the corresponding grayscale figure is intelligible. Color reproduction charges for print are the responsibility of the author. Details of the associated charges can be found on the IEEE Publications page.

Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in these TRANSACTIONS is considered acceptance of this requirement.

*EDICS Webpages:

IEEE TRANSACTIONS ON SIGNAL PROCESSING:
http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/
IEEE TRANSACTIONS ON IMAGE PROCESSING:
http://www.signalprocessingsociety.org/publications/periodicals/image-processing/tip-edics/
IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE / ACM:
http://www.signalprocessingsociety.org/publications/periodicals/taslp/taslp-edics/
IEEE TRANSACTIONS ON INFORMATION, FORENSICS AND SECURITY:
http://www.signalprocessingsociety.org/publications/periodicals/forensics/forensics-edics/
IEEE TRANSACTIONS ON MULTIMEDIA:
http://www.signalprocessingsociety.org/tmm/tmm-edics/
IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING:
http://www.signalprocessingsociety.org/publications/periodicals/tci/tci-edics/
IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS:
http://www.signalprocessingsociety.org/publications/periodicals/tsipn/tsipn-edics/

JOIN SPS TODAY AND RECEIVE

Benefit and Package	Essential Membership	Preferred Membership
Inside Signal Processing eNewsletter	✓	✓
IEEE Signal Processing Magazine	Digital Electronic	Digital Electronic Print
IEEE Signal Processing Content Gazette	✓	✓
Signal Processing Digital Library *Electronic access to seven solely-owned SPS publications through IEEE Xplore®*		✓
SPS Resource Center	✓	✓
IEEE Professional Member Price *Membership through 31 December 2017*	$22.00	$39.00
IEEE Student Member Price *Membership through 31 December 2017*	$11.00	$20.00
Affiliate Member Price *Membership through 31 December 2017*	$96.50	$113.50

In addition, all SPS members receive:

- **Networking** and **collaboration** opportunities with a global network of nearly 17,000 signal processing professionals
- **Discounts** on SPS conferences and workshops, including our flagship conferences ICASSP, ICIP, and GlobalSIP
- **Discounts** on print editions of SPS-sponsored publications
- Eligibility to apply for **travel grants** to SPS conferences
- **Connect** with members near you through local events curated by SPS' 180+ worldwide Chapters
- Career growth and **professional development** tools and resources
- Eligibility to join a **Technical Committee** or **Special Interest Group** to meet SPS members with similar technical interests to develop and strengthen technical communities within signal processing, having a voice in awards, conferences, publications, education, and more
- **Volunteer** opportunities throughout Society activities, including publications, conferences, membership, public visibility, and more.
- **Students** get exclusive access to competitions, job fairs, and networking events

signalprocessingsociety.org

IEEE Signal Processing Society®

IEEE