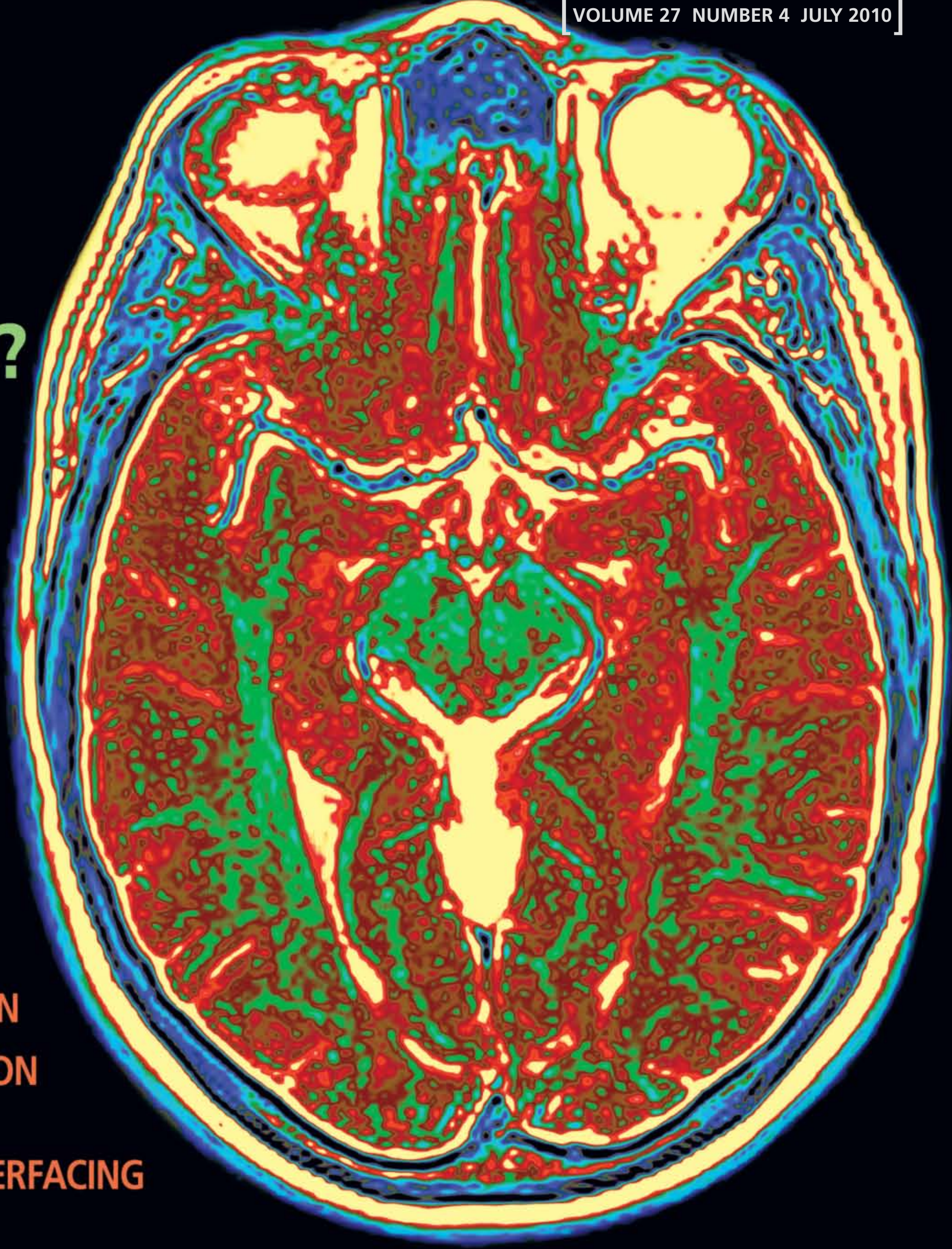
IEEE
Signal Processing
MAGAZINE

VOLUME 27 NUMBER 4 JULY 2010



WHAT'S ON YOUR MIND?

SEE WHAT MEDICAL IMAGING IS ALL ABOUT

PROBING WAVEFORM SYNTHESIS AND RECEIVER FILTER DESIGN

NONRIGID REGISTRATION OF MEDICAL IMAGES

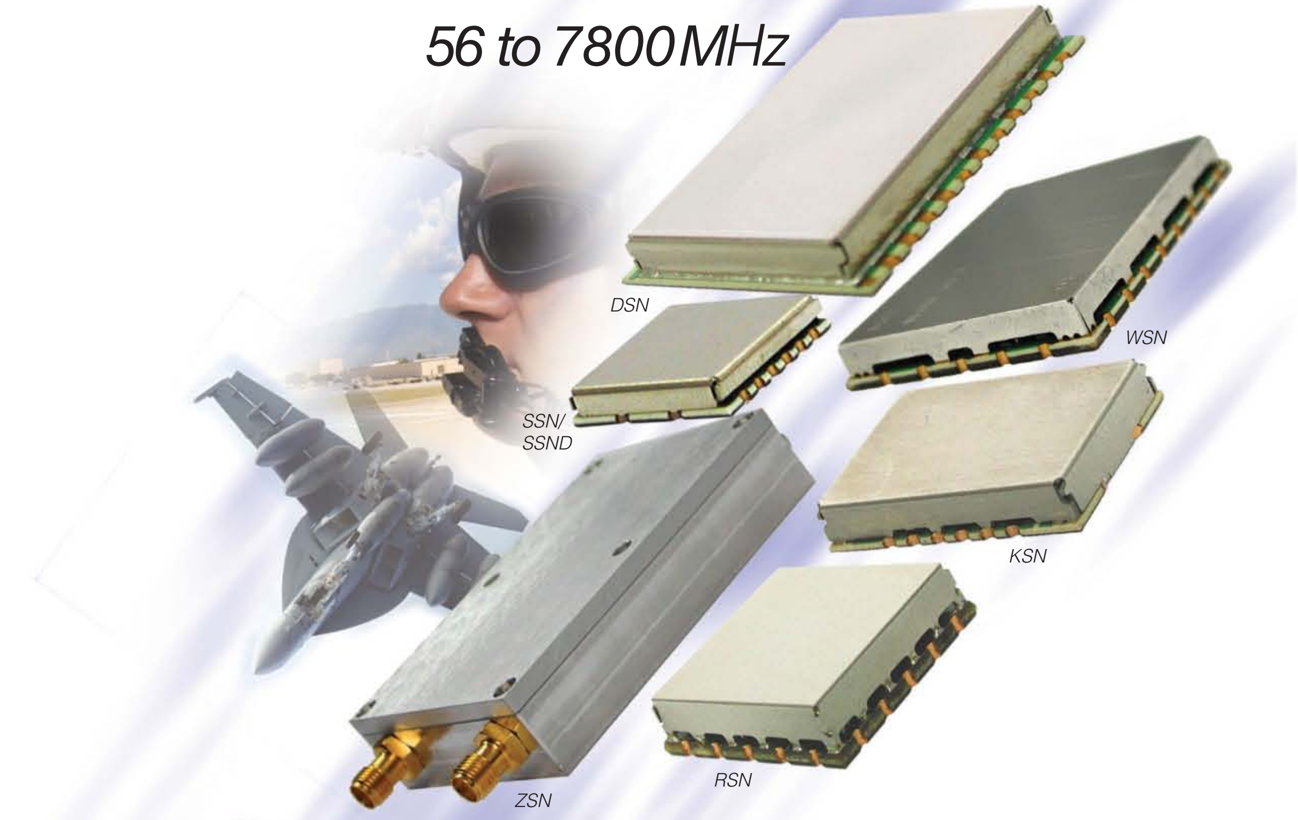
BRAIN-COMPUTER INTERFACING

IEEE Signal Processing Society®

IEEE

OFF-THE-SHELF SYNTHESIZERS

56 to 7800MHz



77 Different Types of Synthesizers from $29.95 qty. 1-9

Offering the needs of a wide range of systems and applications

- Fixed Frequency
- Tunable Narrow Bandwidth
- Tunable Medium Bandwidth
- Tunable Wide Bandwidth
- Tunable Very Wide Bandwidth
- Tunable Fast Settling Time
- Dual Frequency

Photos: Courtesy of U.S. Navy and NASA

Mini-Circuits is an industry leader in synthesizer designs with over 250 standard catalog models and many more custom units designed into major programs supporting a wide range of applications.

TDSCDMA CDMA Repeaters WCDMA EDGE GSM TDMA PCS UMTS L-Band Satellite
Wireless LAN Point-to Point Radio Test Equipment Microwave Radio High Data Transfer Rates

Need a custom model? Please contact our applications department.

We invite you to visit our Website and view our standard models containing comprehensive performance curves, data sheets, PCB layouts, and environmental specifications. ALL catalog models are available for immediate delivery, direct from our web store. Have a unit in your hands for test and integration as early as tomorrow.

Mini-Circuits...Your partners for success since 1969

Mini-Circuits
ISO 9001 ISO 14001 AS 9100 CERTIFIED

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

YONI2™ Patent Pending **The Design Engineers Search Engine** *finds the model you need, Instantly* • For detailed performance specs & shopping online see minicircuits.com

IF/RF MICROWAVE COMPONENTS

477 Rev. A

[VOLUME 27 NUMBER 4]

[CONTENTS]

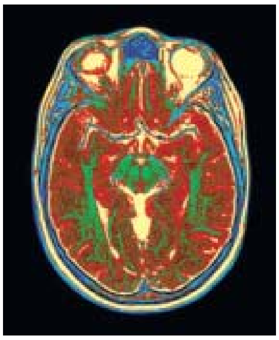
[SPECIAL SECTION—SIGNAL AND IMAGE PROCESSING IN MEDICAL IMAGING]

[FEATURE]

[COLUMNS]

[DEPARTMENT]

[COVER] © PHOTODISC/DON FARRALL



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.

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 $20.00 (first copy only), nonmembers $175.00 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. Copyright©2010 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. 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.2010.936720

[from the EDITOR]

Antonio Ortega
Area Editor, Feature Articles
antonio.ortega@sipi.usc.edu
ahttp://signalprocessingsociety.org/
publications/periodicals.spm

Socializing Digital Signal Processing

First of all, do not be alarmed: this is not about an imminent government takeover of the IEEE Signal Processing Society!

I am, of course, talking about online social networks and their relevance to our community. A quick search through some of the most popular among those, e.g., LinkedIn, Facebook, or Twitter, turns up increasing numbers of our digital signal processing colleagues, sometimes greeted with dismay by the younger set in those communities ("Even my professor is on Facebook!"). Those of you who attended the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) may have participated in the IEEE Thematic Meetings on Signal Processing (THEMES) workshop, which provided evidence that, as a community, we find many interesting research issues to explore in these emerging online systems. But here I'm not talking about what our research can do for online social networks but rather about what online social networks can do for our research.

Just before I headed to Dallas for ICASSP, I had interesting discussions with some colleagues. I learned how, in communications/journalism conferences, participants in this field routinely exchange comments and observations on what is being presented and discussed, immediately, real time, via Twitter. This raised the obvious question: Why not us?

So, with this in mind, at ICASSP I set out to explore our own use of these social networks for research interactions. For this highly nonscientific endeavor I chose Twitter, primarily because of the low effort involved in posting (especially if one does not aim at sending clever and/or informative tweets). It is also easy to search for specific tags (#icassp2010 anyone?), so I was hoping to find out quickly what the buzz was at the conference.

The result of this survey? Including me, probably just a handful of people were tweeting *from* ICASSP *about* ICASSP.

Leave aside the issue of whether any of us should use a specific online social networking tool. The larger issue is what these tools, present and future, mean for our community and our work, and

(continued on page 119)

Digital Object Identifier 10.1109/MSP.2010.937187

Digital Object Identifier 10.1109/MSP.2010.937228

[president's **MESSAGE**]

Mostafa (Mos) Kaveh
2010–2011 SPS President
mos@umn.edu

March Madness—The Good Kind

The raison d'être for each of the IEEE Signal Processing Society's (SPS's) two major conferences, the International Conference on Acoustics, Speech and Signal Processing (ICASSP) and the International Conference on Image Processing (ICIP), is a forum for the exchange of technical information through contributed, invited, and tutorial sessions. However, these meetings, and particularly ICASSP, also serve another important function; they are venues for administrative and planning meetings for the Society's myriad boards and committees. Following are some highlights, decisions, and plans from the administrative meetings at ICASSP in Dallas last March.

The SPS has one of the healthiest financial positions in the IEEE. This good fortune is the results of the popularity and success of our products and services as well as sound management and prudent investment of reserves over many years. It is worth mentioning that while the Society's reserves are strong, only a small portion of those are accessible for our use and there are constraints as to how we can use them. Operational budgets do face occasional challenges, largely because of rules that guide the creation of budgets, as well as potential challenges of new benefits, such as free electronic publications, are reflected on the bottom line. Nevertheless, the Society is utilizing funds allowable by IEEE rules on new initiatives to benefit the members and create greater visibility to the discipline of signal processing. Among the initiatives being planned is an expanded program of tutorials and other learning resources of value for the broad range of our constituency.

Speaking of publications, the submissions and the number of published pages in most of the Society's journals and *IEEE Signal Processing Magazine* continue to grow at astounding rates. This has created significant challenges for the size of some of the print copies and for editorial and production loads. The SPS Publications Board continues to explore avenues for better review and editorial management and tools. For example, to manage the very large volume of submissions, the *IEEE Transactions on Signal Processing* Editorial Board has reorganized into specific technical areas with area editors selected to assist the editor-in-chief. Of course, we are constantly exploring avenues for providing high impact and high visibility publications. *IEEE Signal Processing Magazine* in particular continues to innovate by providing a range of high-quality material and topics of broad interest. For example, in April, *CBS Sunday Morning* did a story about the SETI Institute and radio telescopes. We had an article on that same subject in the March issue of the magazine.

[AMONG THE INITIATIVES BEING PLANNED IS AN EXPANDED PROGRAM OF TUTORIALS AND OTHER LEARNING RESOURCES OF VALUE FOR THE BROAD RANGE OF OUR CONSTITUENCY.]

I had the pleasure of visiting the IEEE Signal Processing Society office last April and had the chance to meet many of the staff who make our publications and other services and operations happen. We are fortunate to have the support of outstanding staff within the Society and from IEEE operations such as the Periodicals Department, so it was great to be able to thank them, in person, on behalf of the signal processing community.

Acting on an initiative by the IEEE Information Theory Society (ITS), the Board of Governors approved a two-year pilot project of closer collaboration between our two Societies through the establishment of mutual liaisons. Prof. Urbashi Mitra is the liaison from the ITS to the SPS; Prof. Nikos Sidiropoulos is the SPS liaison to the ITS. The two liaisons will explore opportunities for joint activities such as sponsorship of workshops and other initiatives in the many areas of joint interest to our Societies.

The Society's Board of Governors ratified the election of the incoming vice president, Awards and Membership (2011–2013), Dr. John Treichler. Further, the current vice president, Awards and Membership, Prof. Mike Zoltowski, appointed the directors of the two new committees under the Membership Board. The Membership Services Committee is directed by Prof. Shuvra Bhattacharyya, and the Industry Relations Committee is directed by Dr. Alex Acero.

I am looking forward, with anticipation, to an exciting ICIP 2010 in September in Hong Kong, along with visits to some of our Chapters and active locations in China before the conference. I will provide more information on these plans in my September 2010 column.

Enjoy the rest of the summer! [SP]

Digital Object Identifier 10.1109/MSP.2010.937230

Ron Schneiderman

DSPs See Gains in Their Impact on New Medical Imaging Designs

Digital signal processing (DSP) is having a major impact on advancing the state of the art of medical imaging.

The advantages of DSP are well established: They operate in real time, they're highly reliable, and they are very energy efficient. They're also relatively inexpensive. But the medical imaging market continues to push for more technical innovation. That's putting more focus on higher image quality and designing smaller systems.

"Over the next few years, we anticipate a significant shift in medical imaging applications from traditional imaging modalities limited to basic diagnostic functions to a new ecosystem comprised of small form factor, highly accurate portable devices," says Susie Inouye, research director and president of Databeans, a semiconductor market research firm.

The rapid development of portable systems has already resulted in handheld, and, in some cases, even wearable, medical and home monitoring devices. DSPs will be pervasive in all of these systems. As a result, medical equipment manufacturers and chip vendors are working hard to expand medical diagnostic applications and introduce new products to a growing market.

General Electric (GE) earlier this year unveiled its vScan machine, which is about the size of a cell phone and sells for under US$10,000. Siemens has upgraded the Acuson P10 handheld scanner it first introduced three years ago.

Toshiba recently entered the portable ultrasound market with a new laptop system. Called Viamo, it's designed mainly for use with immobile patients that need a high-end ultrasound exam. Hitachi also offers a laptop-size system.

GE Sensing & Inspection Technologies has introduced a lightweight (13 lb) and portable digital radiography tool, the DXR250V, that features shorter shot times for minimal radiation exposure in applications that were previously limited to computed radiography or film. The new GE unit can be connected to a laptop to produce images for instant review.

A much smaller firm, Signostics, recently received U.S. Food and Drug Administration (FDA) approval of its cell phone-size Signos Personal Ultrasound system, which weighs about a half a pound.

[DIGITAL SIGNAL PROCESSING IS HAVING A MAJOR IMPACT ON ADVANCING THE STATE OF THE ART OF MEDICAL IMAGING.]

"Signostics overcame some difficult product design challenges in developing its palm-sized ultrasound product," says Patrick O'Doherty, healthcare segment director of Analog Devices, which worked closely with Signostics to provide key signal processing technologies for the data conversion, signal conditioning, and sensing necessary to achieve its design. The Signos covers several medical applications, including abdominal assessments such as bladder, abdominal aortic aneurysm screening, and trauma assessment; musculoskeletal, and basic obstetrics.

SonoSite Inc., another player in the point-of-care market, offers a hand-carried ultrasound system that's used mainly in doctor's offices.

Philips Medical introduced a handheld ultrasound device nearly ten years ago. Called OptiGo, it was taken off the market, reportedly because of doubts at the time about the image quality of a medical imaging device that was so small.

DIFFERENT SYSTEMS, APPLICATIONS

There are several medical imaging technologies.

Magnetic resonance imaging (MRI) offers extraordinarily clear images of the human body and is used to diagnose a wide range of illnesses and injuries. More than 60 million diagnostic MRI procedures are performed worldwide each year.

A noninvasive technique, MRI produces images of the human body without using ionizing radiation. Because of its ability to tailor an exam to meet specific imaging parameters such as the field of view, it is the method of choice to diagnose many different medical conditions, including cancerous tumors, torn ligaments, and Alzheimer's disease.

Computed tomography (CT) is another form of scanning that produces three-dimensional images of internal parts of the body. It's being used increasingly as the technology improves to provide clearer, more detailed pictures for analysis and diagnosis of internal organs, bones, soft tissue, and blood vessels. "Advancements in CT scan imaging will fundamentally change the practice and economics of diagnostic imaging," says Susie Inouye of Databeans. (Today, more than 62 million medical CT scan exams are done in the United States annually, compared to three million in 1980.)

Advancements in systems integration have already helped to significantly boost the number of pictures (or "slice counts") that can be taken using CT machines, improving image detail and quality.

Digital Object Identifier 10.1109/MSP.2010.936828

Digital X-ray is a major step up in diagnostic technology from conventional X-ray systems, where signal degradation from each component consumes more than 60% of the original X-ray signal. By adding a digital detector to digital X-ray imaging, more than 80% of the original image information is captured. The use of digital X-rays also reduces patient radiation dosages and reduces diagnosis time by eliminating photographic processing. High-performance DSPs can control the functions and signal conditioning to acquire and improve the clarity of digital X-ray images. Another key benefit of digital X-ray is its ability to store and transfer the digital images.

Diagnostic ultrasound imaging systems generate and transmit acoustic waves and capture reflections that are then transformed into visual images. The signal processing on the received acoustic waves include interpolation, decimation, data filtering and reconstruction. Programmable DSPs and systems-on-a-chip (SoCs) are designed to implement complex mathematical algorithms in real time to efficiently address all the processing needs of these systems.

Another medical imaging system is positron emission tomography (PET). Like MRI, it is a noninvasive diagnostic technology. It uses radiation emissions from the body (generated by radioactive chemical elements consumed by the patient) to produce physiologic images of specific organs or tissues.

DSPs are normally used in PET systems to handle varying input amplifier gain and to control the photomultiplier tube high-voltage supply and motion control for detector ring assembly and patient entry/exit through the actual system. DSPs can also be used for PET scanner control and signal processing units.

Westside Medical Associates of Los Angeles and Westside Medical Imaging (WMI) of Beverly Hills have recently reported the benefit of early PET scanning to identify Alzheimer's in its early, more treatable phase. "The research investigators at the New York University (NYU) Langone Medical Center have confirmed our long-held belief that we can use advanced imaging for early identification of Alzheimer's disease in patients that have not yet developed symptoms," says Dr. Norman Lepor, professor of medicine at the Geffen School of Medicine at the University of California, Los Angeles and codirector at WMI.

The NYU research team has been using PET with a fluorescent imaging agent called Pittsburgh Compound B that lights up clumps of a protein called beta amyloid that is a characteristic finding of Alzheimer's disease. According to the researchers, not all patients with beta amyloid plaques in their brain develop Alzheimer's.

RADIATION EXPOSURE MAY REQUIRE DEVICE DESIGN CHANGES

Radiation risk has become a big issue in recent months for patients and a hot topic among medical imaging system manufacturers, radiologists, and physicians.

Federal regulators believe CT scans may be necessary to detect a myriad of health issues, but they also detect growing evidence that exposing people to radiation may increase their risk of getting cancer in the future.

So much so that the U.S. FDA Center for Devices and Radiological Health (CDRH) has kicked off a radiation reduction initiative that could force manufacturers of imaging devices to redesign their products so they can alert healthcare professionals when radiation doses exceed recommended levels.

The FDA held the first in a series of conferences in early April to discuss how to protect patients from unnecessary radiation exposures.

The FDA says its goal is to support the benefits associated with medical imaging while minimizing the risks. "The amount of radiation Americans are exposed to from medical imaging has dramatically increased over the past 20 years," says CDRH Director Dr. Jeffrey Shuren. In fact, recent studies indicate the average American's total radiation exposure has nearly doubled in the last three decades, largely due to CT scans and other next-generation imaging tests.

For example, the radiation dose associated with a CT abdomen scan is the same as the dose from approximately 400 chest X-rays. In comparison, a dental X-ray requires approximately one-half the radiation dose of a chest X-ray. The FDA says it intends to issue targeted requirements for manufacturers of CT and fluoroscopic devices to incorporate important safeguards into the design of their machines to develop safer technologies and to provide appropriate training to support safe use by practitioners. The agency held the first in a series of public hearings in late March to solicit input on what requirements to establish.

In a bid to empower patients and increase awareness, the FDA is collaborating with other organizations to develop and disseminate a patient medical imaging history card. This tool, which will be available on the FDA's Web site, will enable patients to track their own medical imaging history and share it with their physicians, especially when it may not be included in their medical records.

The Medical Imaging and Technology Alliance (MITA), an association that represents manufacturers of medical imaging and radiation therapy systems, says it supports initiatives to reduce exposure to unnecessary radiation and minimize medical errors.

The American Society of Radiologic Technologists says it supports MITA's efforts to incorporate a radiation dose check feature on all new CT products, as does the Alliance for Radiation Safety in Pediatric Imaging, which leads the Image Gently campaign to reduce radiation doses for children who undergo medical imaging exams.

Siemens has developed a new imaging system called the Somatom Definition Flash scanner that uses a relatively low dose of radiation and only targets a specific area of the body (see "Radiation Exposure May Require Device Design Changes").

DSP VENDORS SEE GAINS

Several major chip companies are working to advance the state of the art in improving the accuracy and efficiency of medical imaging systems.

Texas Instruments (TI) has long been a leader in providing DSPs and related devices for medical imaging applications and formed a Medical Imaging DSP Group in 2007. The following year, it launched a US$15 million medical university fund to have a "significant effect" in medical technology over the next three to five years.

Ken Nesteroff, TI's DSP medical imaging business development and marketing manager, says ultrasound is one of the better examples of where DSP fit into medical imaging systems (see Figure 1).

"Of course, there are a lot of analog solutions and we build specific parts for that," notes Nesteroff. "On the DSP side, we fit more into the back-end processing. What you typically see is a gigahertz-class DSP in the B-mode, color flow, and Doppler functions, and sometimes the RF demodulation. The back-end function is where you scan-convert the data for display. In a portable system, the industry has moved completely away from a PC back-end to a more system-on-a-chip approach."

TI is currently upgrading the embedded processor software toolkit it introduced in March 2009 to help medical diagnostic ultrasound manufacturers develop more accurate and cost-effective systems, and do it more quickly. Key to the new toolkit, says Nesteroff, will be advances in image processing.

TI also sees opportunities in medical imaging for its newest SoC architecture based on its multicore DSPs that integrates fixed and floating point capabilities. Designed for communications infrastructure equipment, the new DSPs run at up to 1.2 GHz and provide an engine with up to 256 giga multiply-accumulate operations per second (GMACS) and 128 gigaflops (GFLOPS).

Analog Devices, a long-time collaborator to the medical imaging industry, recently introduced a new current-to-digital converter chip that enables high slice count CT systems to capture real-time moving images—such as a beating heart—with a high degree of accuracy and detail. The chip changes photodiode array signals into digital signals and, according to Analog Devices, provides a 50% reduction of CT detection system electronics cost, largely through a more highly integrated design when compared to older models.

"The important thing to remember about any imaging system that is going to be used in medical diagnostics is to maintain image quality with no loss of information that could be discernable to the physician," says Tony Zarola, a strategic marketing manager with the Analog Devices Healthcare Group. Higher image resolution translates into more pixels, and Zarola says that means more data and higher demands on back-end image processing.

An obvious goal is to reduce the exposure (less scan time) to harmful X-ray images while obtaining more information during the scan. In terms of what this means for the electronics in the system, more scan lines means more channels, higher image resolution translates to more pixels, and a higher signal-to-noise radio provides less noise and therefore better contrast.

[THE USE OF DIGITAL X-RAYS ALSO REDUCES PATIENT RADIATION DOSAGES AND REDUCES DIAGNOSIS TIME BY ELIMINATING PHOTOGRAPHIC PROCESSING.]

"More data being transmitted from the receivers, increasing the channel count requires increased bandwidths across the system," adds Zarola. "This can cause challenges with transfer of data over existing infrastructures, which are limited in bandwidth."

The benefits of DSPs, he says, are significant, ranging from a reduction in bandwidth to the use of smart compression algorithms. (Lossy compression could be used, but then the resulting impact on image integrity would need to be characterized.) "For better image quality, various post-processing image enhancement algorithms that improve contrast or reduce the effects of system noise could be employed," says Zaroloa. "Again, the challenge would be to keep the image integrity."

A HUGE MARKET

Medical imaging is already a huge market and it continues to grow, largely due to advancements in the technology, and growing popularity of portable and hand-carried imaging products. The global market for medical imaging devices is projected to reach about US$37 billion by 2015, according to a market study by Reportlinker.

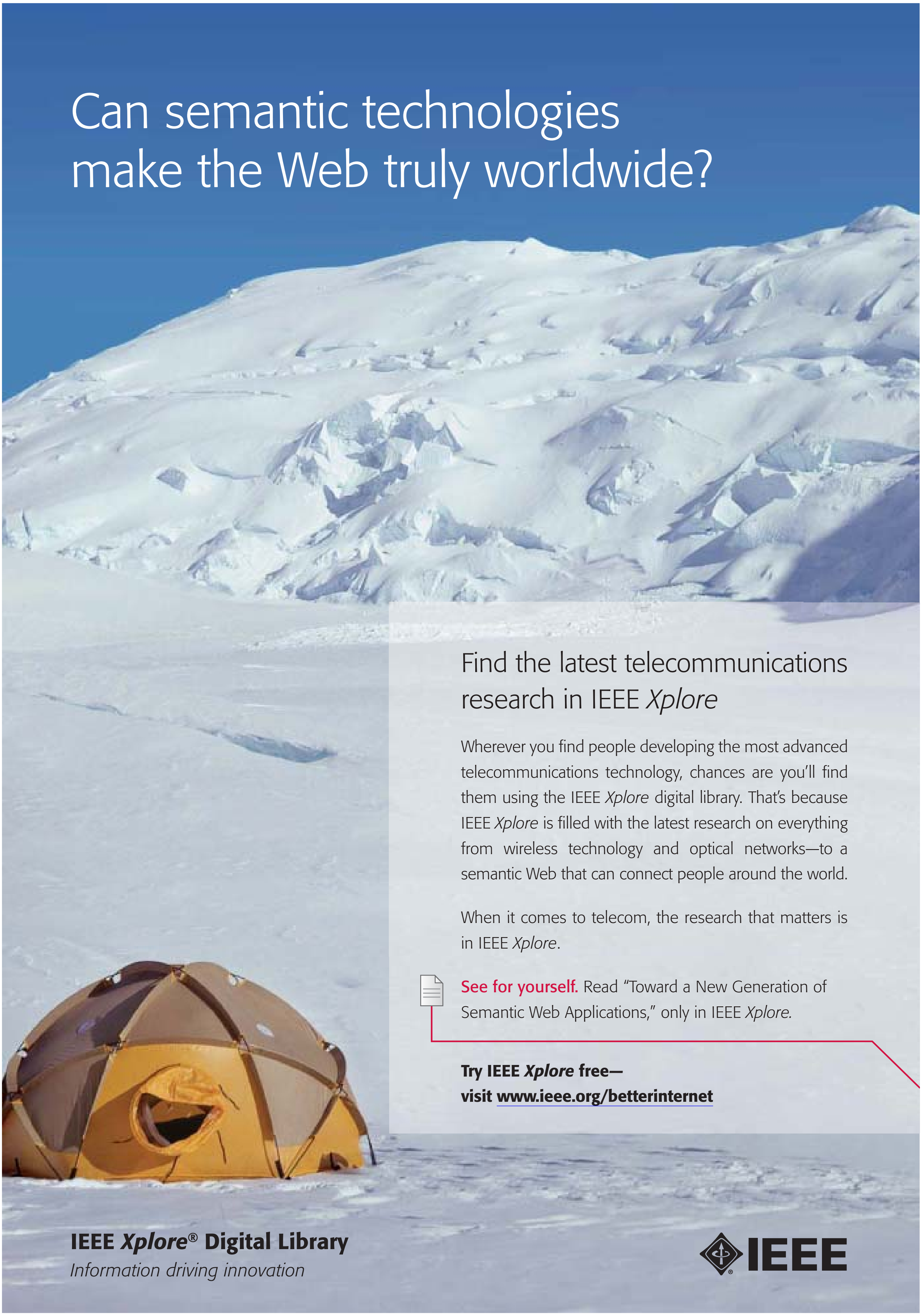
MRI is expected it be the fastest growing imaging modality with a compound annual growth rate (CAGR) of 9.8% during the period 2005–2015.

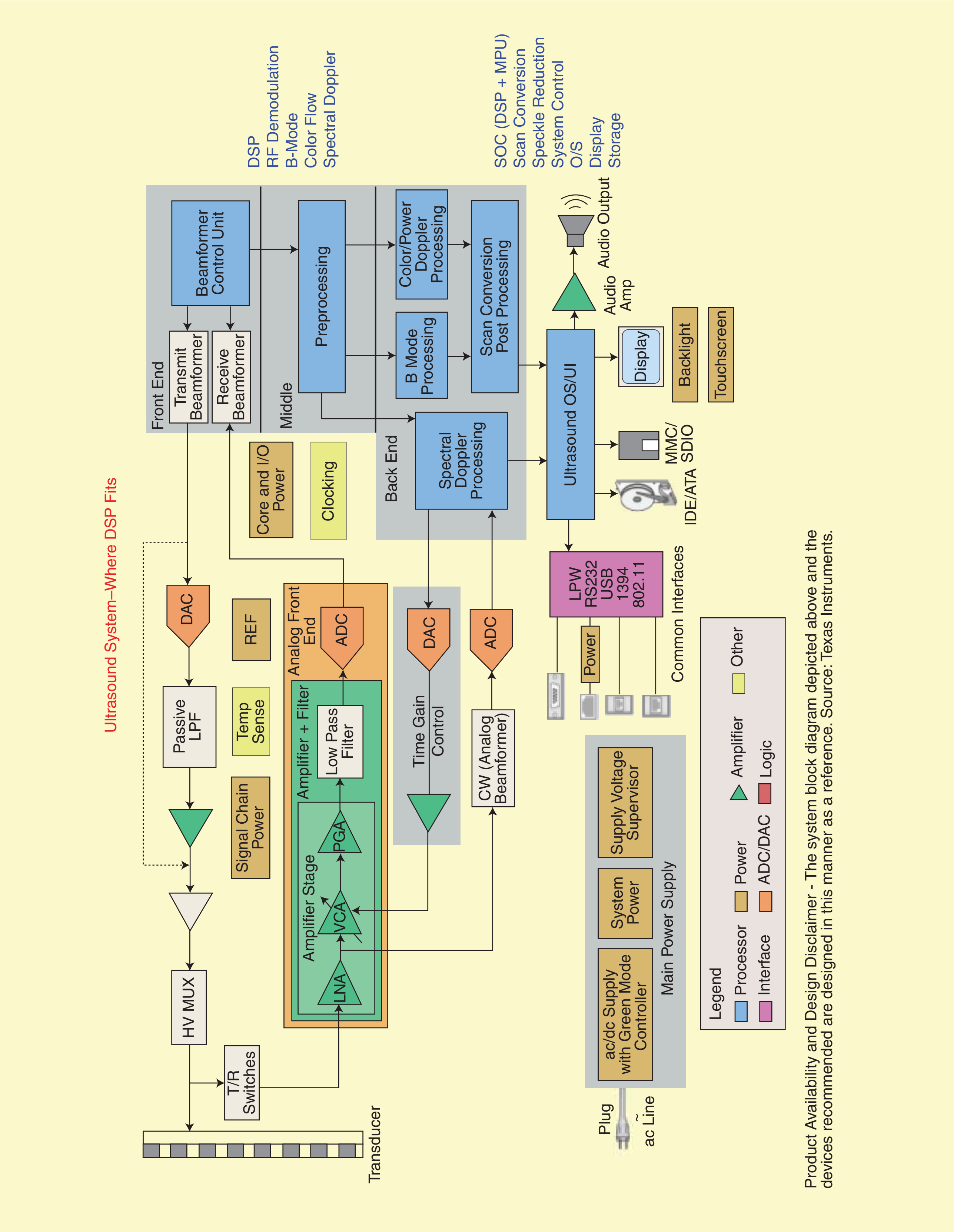
Another market research group, Global Industry Analysts, says the U.S., Japan, and Europe account for more than 85% of the world market installed base of CT scanners. According to Global Industry Analysts, the global CT scanner market is dominated by four companies: GE Healthcare, Siemens Healthcare, Toshiba Medical Systems, and Philips Healthcare. Other major players include Hitachi Medical Corp. and Shimadzu Medical Systems.

Rapid upgrades in technology have also had an impact on CT scanners. A key trend in the CT segment is the shift towards combination scanners, which are primarily hybrid scanners comprising PET and CT imaging capabilities.

Global Industry Analysts says ultrasound has won a growing share of the medical imaging market since its introduction in the early 1950s. The miniaturization of ultrasound devices and continued incorporation of system electronics into ultrasound technology is a major trend and accounts for much of the success of this imaging technology.

The overall market for ultrasound equipment is near saturation levels in the United States; however, cardiology continues to represent a fast growing end-use segment of ultrasound with revenues in the United States projected to reach US$684 million in 2010. This market is essentially driven by the need for replacing and upgrading aging equipment with new, more technically advanced, systems. The U.S. and Europe collectively account for about 60% of the global medical ultrasound equipment market, although the Asia-Pacific markets are growing rapidly according to the market research firm.





[FIG1] This system block diagram is a reference design for DSP and other devices that Texas Instruments suggests can be used in an ultrasound medical imaging system design. (Figure used with permission.)

(continued on page 134)

[from the GUEST EDITORS]

Miles N. Wernick, Charles A. Bouman,
Richard M. Leahy, and James S. Duncan

The Roles of Signal Processing in Medical Imaging

Exploratory surgery, an approach in which the physician looks directly within the patient for the source of an ailment, was once the principal method of visualizing disease. While this approach still has its place, it has been largely supplanted by medical imaging, a vast arsenal of technologies capable of producing detailed and highly informative images of the body's internal structure and function.

Medical imaging uses a wide variety of physical phenomena, ranging from x-ray attenuation to acoustic wave propagation, to measure a staggering number of variables relating to the patient's health. The earliest medical images showed only structural information, such as the appearance of bones; however, many modern techniques can now evaluate intricate biological processes, such as metabolism, distribution of chemical receptors, abnormal heart motion, or deposition of amyloid plaque associated with Alzheimer's disease.

Thus, medical imaging is used not only to diagnose disease; it also provides an essential tool for understanding human biology and is widely used to evaluate the effectiveness of new drugs. In many instances, medical imaging is used to plan surgical procedures and even to guide these procedures while in progress. For example, in this issue, the article by Mountney et al. describes how image analysis can be used as an aid in robotic-assisted minimally invasive surgery.

THE ROLE OF SIGNAL AND IMAGE PROCESSING

In the world of medical imaging, signal and image processing are involved in every stage of the process. Most types of medical images are computed as the solution of a complicated inverse problem. In some cases, the data from which the inverse problem is solved are themselves derived from significant signal-processing steps. Once obtained, medical images are often analyzed and interpreted automatically by sophisticated image-processing and machine-learning techniques. In addition, signal processing is an essential tool used in the design and evaluation of imaging devices, and in the assessment and prediction of diagnostic performance.

[MEDICAL IMAGING IS USED NOT ONLY TO DIAGNOSE DISEASE, IT ALSO PROVIDES AN ESSENTIAL TOOL FOR UNDERSTANDING HUMAN BIOLOGY.]

The general public—and indeed many of our colleagues in the signal-processing field—are unaware of the degree to which signal and image processing play an essential and enabling role in medical imaging technology. And the importance of signal processing continues to grow rapidly as the technology continues to mature and advance, and as the medical field continues to be more accepting of the role of computers and technology in clinical practice.

In 1997, owing to rapid developments in the field, *IEEE Signal Processing Magazine* published a special issue devoted to medical imaging in which several of the basic types of imaging were introduced to the broader signal processing community. Inspired by recent advances, we now revisit the topic, this time focusing on the diverse and expanding roles that image processing plays in this important field.

OVERVIEW OF THIS SPECIAL ISSUE

Roughly speaking, the articles in this special issue are divided into two main groups. The first four articles discuss various ways that medical images are analyzed automatically by a computer. The last three articles describe how mathematical techniques contribute to formation of the images themselves.

The first article, by Mountney et al., describes how image analysis can be used in real time to guide minimally invasive surgery performed with robotic assistance. In this work, a three-dimensional description of the patient's tissue (which is soft and constantly deforming during surgery) is determined and tracked from stereo images taken with a laparoscope camera threaded through a small incision in the patient.

The second article, by Wernick et al., illustrates the growing and varied roles of machine-learning techniques in medical imaging, providing examples relating to computer-aided diagnosis, content-based image retrieval, prediction of diagnostic accuracy, and, finally, the rapidly growing field of functional brain mapping, which is providing an unprecedented window into the workings of the human mind.

This leads directly into the next two articles, both of which pertain specifically to study of the brain. The third article in this issue, by Correa et al., describes mathematical techniques for integrating information gathered from a wide array of brain imaging techniques, such as functional magnetic resonance imaging and electroencephalography.

(continued on page 140)

Digital Object Identifier 10.1109/MSP.2010.936727

IEEE Signal Processing Society
14th DSP Workshop & 6th SPE Workshop
Enchantment Resort , Sedona, Arizona
January 4-7, 2011
www.dspe2011.org

Organizing Committee

General Chairs
Lina Karam, Arizona State University
Ronald Schafer, Hewlett-Packard Labs

DSP Technical Program Chairs
James McClellan, Georgia Tech
Ali Sayed, UCLA

SPE Technical Program Chairs
Gail Rosen, Drexel University
Thad Welch, Boise State University

Advisory Committee
Ahsan Aziz, National Instruments
Khaled El-Maleh, Qualcomm
Gene Frantz, Texas Instruments
Loren Shure, Mathworks
Mark J.T. Smith, Purdue
Martin Vetterli, EPFL

Finance
David Frakes, Arizona State University

Publicity
Andreas Spanias, Arizona State University

Social Programs
Cathy Wicks, Texas Instruments

International Liaisons
Julien Epps, UNSW, Australia
Ramón Rodríguez Dagnino, ITESM, Mexico
Hideaki Sakai, Kyoto University, Japan
Abdelhak Zoubir, Darmstadt Univ., Germany

Call for Papers

The 2011 IEEE Digital Signal Processing (DSP) Workshop and IEEE Signal Processing Education (SPE) Workshop will be held jointly January 4 to 7, 2011, at the award-winning Enchantment Resort. The Enchantment Resort is located 5 miles from Sedona in Boynton Canyon, two hours north of the Phoenix / Scottsdale metropolitan area and two and a half hours south of the Grand Canyon. The venue is surrounded by the Coconino National Forest and Red Rock Secret Mountain Wilderness. The area is revered by the Apache Native Americans as the birthplace of their tribe and holds ancient ruins of Native American cliff dwellings.

The goals of the workshops are to bring together leading engineers, researchers, and educators in signal processing from around the world to discuss novel signal processing theories, methods, and applications. The DSP/SPE Workshops will feature prominent plenary speakers from the signal processing community as well as technical sessions for presenting contributed papers.

Topics for the DSP Workshop include, but are not limited to:

- Sampling, extrapolation, and interpolation
- System modeling, representation, and identification; deconvolution
- Filtering and adaptive systems
- Stationary signals and spectral analysis
- Non-stationary signals and time-frequency analysis
- Multi-rate signal processing and wavelets
- Detection, estimation, and classification
- Signal enhancement, restoration, and reconstruction
- Nonlinear signal processing
- Multi-dimensional signal processing; image and video processing
- Implementations of Signal Processing Systems
- Distributed signal processing
- New directions and applications

Topics for the SPE Workshop include, but are not limited to:

- Signal processing education in non-traditional venues
- Novel laboratory, computer-based, and distance teaching methods
- Signal processing across the engineering curriculum
- DSP curriculum issues (early/late, simulation/real-time, theory/practice)
- DSP outreach issues

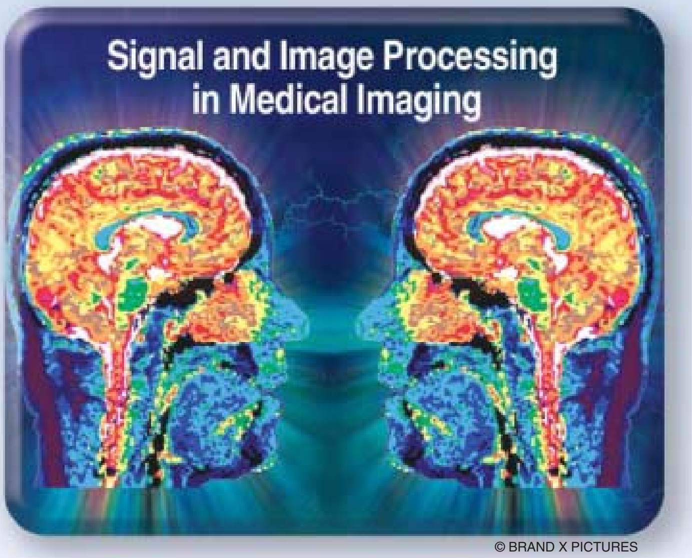
Paper Submission: Prospective authors are invited to submit double-column papers of no more than six (6) pages including title, authors' names and contact, abstract, introduction, background, proposed method, results, figures, and references. Submission instructions and templates for the required paper format are available at www.dspe2011.org.

Important Deadlines:
Submission of Papers: August 30, 2010
Notification of Acceptance: October 11, 2010
Authors' Registration Deadline: October 25, 2010
Submission of Accepted Camera-Ready Papers: November 8, 2010.
Advance Registration and Resort Reservation Deadline: November 15, 2010

[Peter Mountney, Danail Stoyanov, and Guang-Zhong Yang]

Three-Dimensional Tissue Deformation Recovery and Tracking

[Introducing techniques based on laparoscopic or endoscopic images]



© BRAND X PICTURES

Recent advances in surgical robotics have provided a platform for extending the current capabilities of minimally invasive surgery by incorporating both preoperative and intraoperative imaging data. In this tutorial article, we introduce techniques for in vivo three-dimensional (3-D) tissue deformation recovery and tracking based on laparoscopic or endoscopic images. These optically based techniques provide a unique opportunity for recovering surface deformation of the soft tissue without the need of additional instrumentation. They can therefore be easily incorporated into the existing surgical workflow. Technically, the problem formulation is challenging due to nonrigid deformation of the tissue and instrument interaction. Current approaches and future research directions in terms of intraoperative planning and adaptive surgical navigation are explained in detail.

Digital Object Identifier 10.1109/MSP.2010.936728

INTRODUCTION

Over the past two decades, technological innovations have played a major role in reshaping the general practice of surgery. Solid-state cameras and fiber optic devices have made minimally invasive surgery (MIS) a reality. In MIS, specialized instruments are inserted into the anatomy through small access ports and operated under remote video guidance. By avoiding large incisions, MIS greatly reduces patient trauma, postoperative recovery period, and the risk of comorbidity. These advantages have made MIS a viable treatment option for a wider range of patients [1], [2].

Recently, robotic technologies have been used to overcome the limitations of traditional MIS tools and provide the control and maneuverability required for precise microsurgical tasks. Robotic devices represent one of the most promising enhancements in modern operating theatres for MIS. They facilitate the performance of dexterity demanding procedures with improved repeatability and precision through the use of microprocessor controlled mechanical wrists. By using master-slave setups, surgical robots have been shown to significantly improve the

1053-5888/10/$26.00©2010IEEE

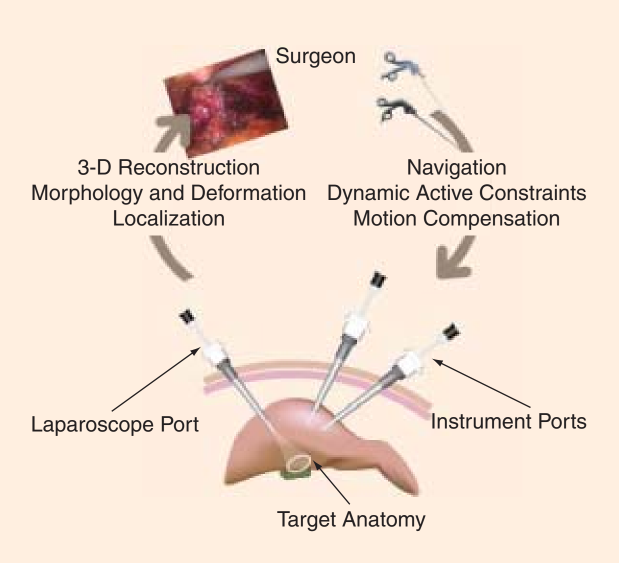
ergonomics in the operating theatre, enable the use of motion scaling and provide a unique platform for real-time navigation based on multimodal patient specific imaging and sensing data.

For performing complex procedures using robotic-assisted MIS, medical image computing plays an important role for improving the surgeon's operating capabilities. Despite the advantages of robotic-assisted MIS instruments, performing microsurgical tasks in a highly dynamic environment is challenging. This is reflected in complex procedures such as robotic-assisted, beating heart totally endoscopic coronary artery bypass (TECAB) surgery, for which, despite the apparent patient benefits, clinical uptake has been slow [3]. While imaging modalities such as intraoperative magnetic resonance imaging and computed tomography can provide accurate information about the tissue morphology, they are constrained by the operating environment mainly due to their size and accessibility. Optical techniques based on laparoscopic or endoscopic cameras provide a unique opportunity for recovering the morphology, as well as the structure of the soft tissue in situ. In MIS, recovering tissue deformation is essential for coregistering intraoperative and preoperative data. It is also important for providing intraoperative guidance and accurately fusing multimodality intraoperative information. With robotic assistance, the recovered tissue deformation can further be used for providing motion stabilization and prescribing dynamic active constraints to avoid critical anatomical structures such as nerves and blood vessels as illustrated in Figure 1.

In this tutorial article, we provide an explanation of the physical configuration of the optical imaging environment in MIS with a geometric camera model and camera calibration. This serves as the basis of techniques for recovering 3-D soft-tissue deformation and relative pose of the laparoscopic cameras. We describe how these techniques can be used for tissue deformation tracking and 3-D reconstruction, with specific focus on the use of a moving camera model for structure recovery. Quantitative validation is discussed to highlight the practical challenges involved for in vivo applications. To summarize we discuss the major challenges and future research directions, particularly in dealing with deformable tissue structures.

OPTICAL SETUP

The laparoscope camera used in MIS is typically inserted into the patient via a small incision or natural orifice. The surgeon maneuvers the external, proximal end of the laparoscope to navigate through the body via a video displays. The MIS environment is illuminated with a light source embedded in the laparoscope. Figure 2 shows the optical configuration of several laparoscopes and example images displayed to the surgeon. Quantitative measurements can be made from laparoscopic images only if the instrument has been accurately modeled and calibrated.

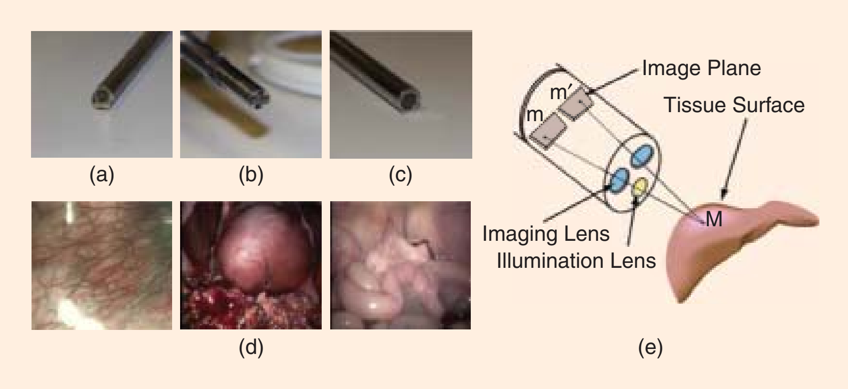


[FIG1] A schematic diagram showing the information flow in robotic-assisted MIS. By using information from the laparoscopic cameras, it is possible to recover tissue deformation in 3-D, which permits intraoperative navigation, motion compensation and dynamic active constraints.

The camera of a laparoscope can be modeled by its optical characteristics called intrinsic parameters and its position and orientation in a world coordinate system called extrinsic parameters [4]. Typically, the pinhole projection model is used to describe the mapping of a 3-D point $\mathbf{M} = [X \ \ Y \ \ Z \ \ 1]^T$ in homogeneous coordinates onto the image point $\mathbf{m} = [x \ \ y \ 1]^T$ as a matrix multiplication

$$\mathbf{m} = \mathbf{K}\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix}\mathbf{M} = \mathbf{PM}, \qquad (1)$$

where $\mathbf{K}$ is a matrix of the intrinsic camera parameters and $\mathbf{R}$ and $\mathbf{t}$ describe the extrinsic orientation of the device in the world coordinate system. Figure 2(e) shows a schematic illustration of this model in a stereo configuration. Lens distortion can be effectively modeled using radial and tangential distortion coefficients [5], [6].



[FIG2] (a) A 30° laparoscope, (b) a stereo laparoscope with two point light sources, (c) a 0° laparoscope with circular light source, (d) example images acquired during MIS, and (e) schematic of a laparoscope with imaging optics observing a sample of tissue in 3-D.

[TABLE 1] SUMMARY OF METHODS USED FOR 3-D RECONSTRUCTION FROM IMAGES IN MIS.

SFS ASSUMPTIONS	STEREO APPROACHES	ACTIVE TECHNIQUE
ORTHOGRAPHIC [10]	COMPUTATIONAL [11], [12]	FIDUCIAL [13], [14]
PERSPECTIVE [6,] [15], [16]	SURFACE PRIORS [17]–[19], [21]	ONE SHOT [20], [22]
ILLUMINATION [23]	CUE FUSION [6], [24]	PROGRESSIVE [2], [25], [26]

In general, the unknown parameters of the laparoscope model are estimated by a preoperative calibration process. To obtain these unknown parameters, certain constraints are usually imposed on the projection of points, with known coordinates in the 3-D world, onto the image plane. There are several well-established algorithms for this procedure from the computer vision communities and implementations of these methods are available online [7].

After calibration, the metric 3-D structure of the surgical scene can be recovered given the correspondence of image primitives [m and m' in Figure 2(e)] among multiple views of the surgical site. This process is called triangulation [4], which is also illustrated in Figure 2(e).

RECOVERING SOFT TISSUE 3-D SHAPE

Recovering 3-D information from images is a long-standing problem in computer vision. Typical solutions are stimulated by our basic understanding of biological vision systems and the intrinsic relationship of how 2-D images are acquired from 3-D space. The early work of Marr [8] led to the establishment of shape-from-X, where different visual cues can be used to infer information about the shape and position of objects with respect to the camera. The wealth of research in this area has resulted in many publications [9]. In this section, we will only summarize those approaches reported in MIS.

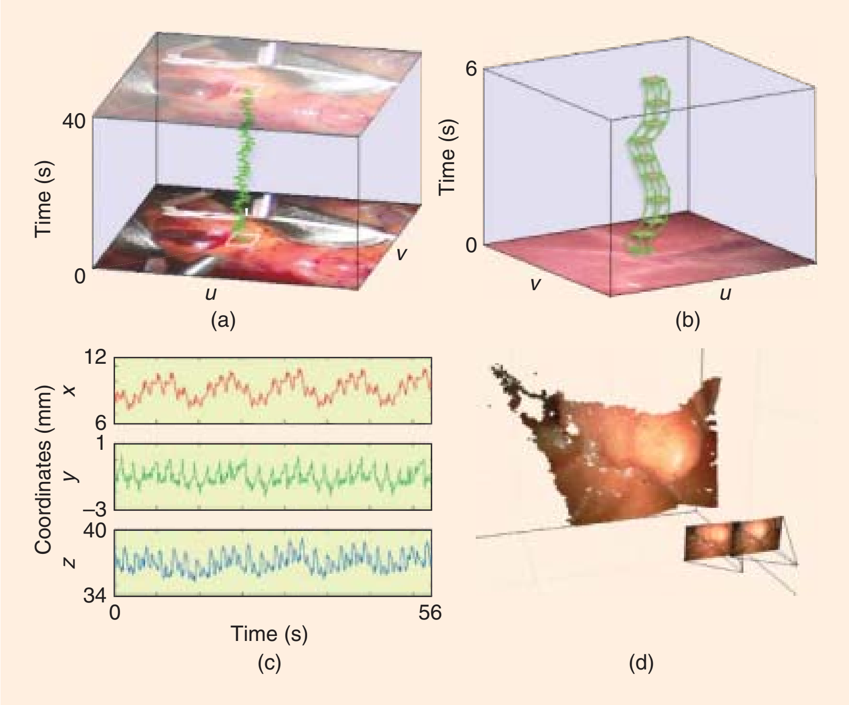
Approaches to 3-D tissue surface reconstruction are summarized in Table 1 and an example is shown in Figure 3(d). They can be broadly divided into passive and active techniques. Passive techniques do not introduce additional light or sensing devices into the MIS environment and are purely based on the existing images as observed by the operating surgeon. The two main visual cues that have been exploited are shading and stereo.

For shape-from-shading (SFS), laparoscopic images do not obey many of the traditional assumptions used to simplify the bidirectional reflectance distribution function (BRDF). Lambertian reflectance is not compatible with specular reflections, which are common due to the mucus layer of the soft tissue and the relatively high intensity of the laparoscopic light source. Furthermore, the assumption of a light source located at infinity is not satisfied due to the copositioning of the light source at the tip of the laparoscope. In addition, the camera cannot be assumed to perform orthographic projection as perspective effects and lens distortions are significant in laparoscopic images.

Therefore, the special optical arrangement between the scope, illumination source, and the surgical scene must be used to simplify the image irradiance equation. This was first proposed by Rashid and Berger in 1992 [10] where the light source and the optical centre of the camera were considered to be coincident. This approach was subsequently combined with the assumption that the BRDF is a monotonically decreasing function with respect to the viewing angle [21]. More recent work has expanded the camera projection model to incorporate lens distortion [15] and perspective projection [6], [16]. The assumption of coincident camera and light source positions has also been relaxed [23], although this requires the calibration of the relative positions [27].

One of the main drawbacks of SFS approaches in MIS is that the information recovered is not in a metric coordinate space and only relative surface orientation information can be measured. Passive stereo techniques and SFS are complementary and can be combined to overcome this limitation [24].

Early work on stereo methods in MIS used a simple normalized cross-correlation algorithm [11]. Subsequently this was adapted to incorporate hierarchical



[FIG3] (a) A region tracked on the cardiac surface illustrating motion from the cardiac and respiratory cycles, (b) a region tracked on the liver illustrating motion resulting from respiration, (c) the tracked 3-D motion of a region on the surface of the heart, and (d) a dense stereo reconstruction of the tissue surface.

solutions with a geometric surface prior and to recover the 3-D shape of the heart [17]–[19]. The use of explicit assumptions (e.g., smoothness) about the observed soft-tissue surfaces enables the reconstruction of homogenous tissue regions but does not handle discontinuities arising at instrument-tissue boundaries. To address this issue, methods based on a sparse set of salient features have been used to first recover a sparse 3-D reconstruction of the surgical site and then propagate this information to achieve a semidense 3-D map [28].

It is worth noting that an important feature of MIS images is the abundance of specular reflections. They are view dependent and can cause significant errors in recovering 3-D structure and tracking deformation. It is therefore necessary to correctly identify these regions prior to stereo correspondence [19], [29]. Alternatively they can be used as constraints when the illumination direction is known or as a starting point in SFS algorithms [24], [27].

The main limitation of passive reconstruction techniques is that they have limited robustness when dealing with the dynamic environment of MIS. For this reason, methods based on fiducial markers or the use of structured lighting have been proposed. Fiducial markers are predominantly used for temporal tissue tracking, which is discussed in more detail in the following section. In terms of structured lighting, an overview of the general techniques is provided in [30]. In surgery, the use of light projection for 3-D measurements has attracted extensive attention [25]. For augmented reality (AR), a structured light system was developed to recover the shape of the surgical site [22]. Subsequently, methods based on a laser plane sweeping over the surgical scene have been developed [2], [26]. All of these systems require an additional instrument port, which has not been clinically popular.

More recently, the use of projected coded patterns has been investigated [20] and methods based on time-of-flight technologies have been explored. They have been shown to produce promising results, albeit at limited resolution and frame rates with the current technologies [31].

SOFT-TISSUE TRACKING AND MORPHOLOGY ESTIMATION

TISSUE TRACKING

The 2-D/3-D morphology and dynamic motion of soft tissue can be recovered by temporally tracking regions of interest in the image. This approach is illustrated in Figure 3(a) and (b) and has been used to recover 3-D tissue morphology and deformation in a variety of anatomical regions as summarized in Table 2. The problem of locating a region of interest in one image and finding the corresponding region in another is difficult in MIS. This is because MIS images can be low in contrast, noisy, and poorly illuminated. The appearance of tissue also varies greatly from homogenous, to highly textured and many regions contain view-dependent specular reflections. It is also necessary to deal with occlusion by surgical instruments, image artifacts, and dynamic effects such as bleeding and cauterization smoke. The performance of a region-tracking algorithm is largely influenced by how distinguishable the region is from its surroundings. This is affected by what regions are detected for tracking, how the region is represented in image or feature space, and the matching strategy used to locate the corresponding region in a new image or video frame.

Region detection is the process of identifying salient regions in the image that are distinguished from their surroundings. Passive techniques that detect naturally occurring features such as vessels, corners, or blobs [32]–[37] are preferred as they do not interfere with the surgeon's view or require user interaction. A comparison of region detectors in MIS is provided in [38]. Tissue can appear homogenous, making region detection challenging. This can be overcome by manually selecting regions [39], [40], using fiducial markers [13], [14], or by marking the tissue of interest (e.g., with diathermy) [39], [41]. These active approaches limit the number of tracked regions.

In general, the region can be represented in **image space** or feature space. In image space, the region is simply represented by pixels as an image patch or template [33], [39]. The main problems with this approach are that the representation is not invariant to large image transformation and the image information may not be sufficient to distinguish a region from its surroundings. Alternatively, descriptors can be used to represent the region in feature space. Feature descriptors select what information from the image will be used (e.g., gray scale, color, and gradient) and how this information will be represented (e.g., energy in the cooccurrence matrix [40], nonuniformity of the run-length matrix [40], probability density histograms [41], histograms of gradients [34], contours, and active appearance models).

Descriptors can be made invariant to image transformation such as scale and rotation through explicit modeling. However, ad hoc modeling of nonlinear deformation is not trivial. Selecting a feature descriptor is context specific and the performance of descriptors can be affected by low-contrast images changes in illumination and specular highlights, making the selection of a robust descriptor challenging. In [13], [34], and [40], machine-learning techniques are used to select and combine discriminant descriptors.

[TABLE 2] SUMMARY OF TISSUE MORPHOLOGY AND STRUCTURE ESTIMATION METHODS APPLIED IN MIS.

	RECOVERED SCENE GEOMETRY	
ORGAN	STATIC	DEFORMING
HEART	[11], [58], [67]	[13], [14], [17], [18], [29], [32], [33], [35], [36], [40], [43], [49], [51], [76]
ABDOMEN / LIVER / GALL-BLADDER / KIDNEY	[37], [59], [64], [69]–[71], 74]	[34], [35], [39], [41]
COLON	[53], [55], [57]	–
BLADDER	[60]–[63]	–
ESOPHAGUS	[54], [56], [72]	–
SINUS	[73]	–

For tracking purposes, the region representation can be created on the first frame and remain constant or updated at each frame. Updating enables temporal persistency to be assumed but can lead to error propagation.

Matching strategies can be categorized as recursive methods or "tracking by detection" [42]. Recursive methods such as Lucas Kanade (LK) attempt to minimize the difference between the region representation and a region in the new image. LK operates in image space and uses the previous location of the region to search for a match locally. This minimization approach works well on small deformations and has been successfully applied to MIS [17], [29], [32], [33], [36], [39], [43]. However, recursive approaches using image space can be sensitive to changes in illumination and specular highlights. They are not well suited to dealing with occlusion and require frame-to-frame updates, leading to error propagation.

In tracking by detection, the region detector is applied to each new video frame to extract a set of potential matches. This set is searched to find a match by comparing feature descriptors. Matching strategies can be one to one (e.g., nearest neighbor), one to many (e.g., nearest neighbor ratio) or many to many (e.g., random sample consensus (RANSAC) [44]). Detectors and descriptors can be complementary such as SIFT [45] and SURF [46]. Tracking by detection is well suited to dealing with occlusion as no temporal information about the region's location is required. The main problem with the application of these techniques in MIS is related to the ad hoc assumptions they make about what image features to use and the expected image transformations. In addition, this approach is dependent on the region detector to correctly locate the region in each new video frame and the global uniqueness of the region as represented in the feature space. Tracking by detection has been applied in MIS [34], and in [35], an approach is proposed which exploits a recursive technique (which requires no prior knowledge) to learn a feature descriptor online.

TISSUE MORPHOLOGY MODELING

Extracting and modeling the 2-D/3-D motion of dynamic tissue is an important prerequisite of image-guided surgery. The 3-D position of tissue, shown in Figure 3(c) can be estimated with a stereo laparoscope as described earlier or with a monocular laparoscope based on fiducial markers with known geometry [13]. In practice, tissue deformation can be caused by the cardiac and respiratory cycles, tissue tool interaction, or muscular contraction.

Deformation resulting from cardiac and respiratory cycles can be modeled as quasi-periodic or periodic signals [47]. Respiration during MIS is usually regulated by a ventilator, creating an asymmetric periodic signal with an extended exhale phase. For example, the effect of respiration on the liver is modeled in [41] by a prototype repetitive controller and using a weighted-frequency Fourier linear combiner in [48]. The motion of the cardiac surface, however, is more complex as it contains deformations caused by both the cardiac and respiratory cycle. The deformations can be decoupled [33], [35] into their intrinsic components or considered together. A number of approaches have been suggested for modeling cardiac motion, which include Fourier series [49], vector autoregressive models [49], Taken's theorem [36], and linear parameter variant finite impulse response models [50]. Information from the ventilator and electrocardiogram (ECG) has also been incorporated to increase accuracy [51]. Modeling large-scale, nonperiodic tissue deformation caused by tissue-tool interaction or muscular contraction is more challenging. It is likely to require the application of statistical shape, finite element, and biomechanical models such as those used in needle steering and surgical simulators [52].

STRUCTURE AND CAMERA MOTION ESTIMATION

The methods described in the previous sections are based on the assumption that the laparoscopic camera is static. This is not true in practice, particularly with the recent emergence of natural orifice transluminal endoscopic surgery (NOTES) or single port access (SPA) techniques. In this section, we will describe two approaches for recovering the structure of the MIS environment, as well as the camera position: structure from motion and simultaneous localization and mapping (SLAM). These competing techniques are compared schematically in Figure 4. Both approaches are based on the assumption that the structure of the environment is relatively stable. It is worth noting that this is a strong assumption for MIS. Nevertheless, these methods have been applied to various parts of the anatomy (Table 2) where tissue motion or deformation is minimal. The extension of these techniques to nonstatic environments will be discussed.

STRUCTURE FROM MOTION

Structure from motion [4] is a computer vision technique developed to recover the structure of a scene and the motion of the camera. A wide variety of approaches exist. However, the basic framework contains three components as illustrated in Figure 4: 1) image registration and frame-to-frame camera motion estimation; 2) global optimization or bundle adjustment where multiple images are registered; and 3) scene reconstruction.

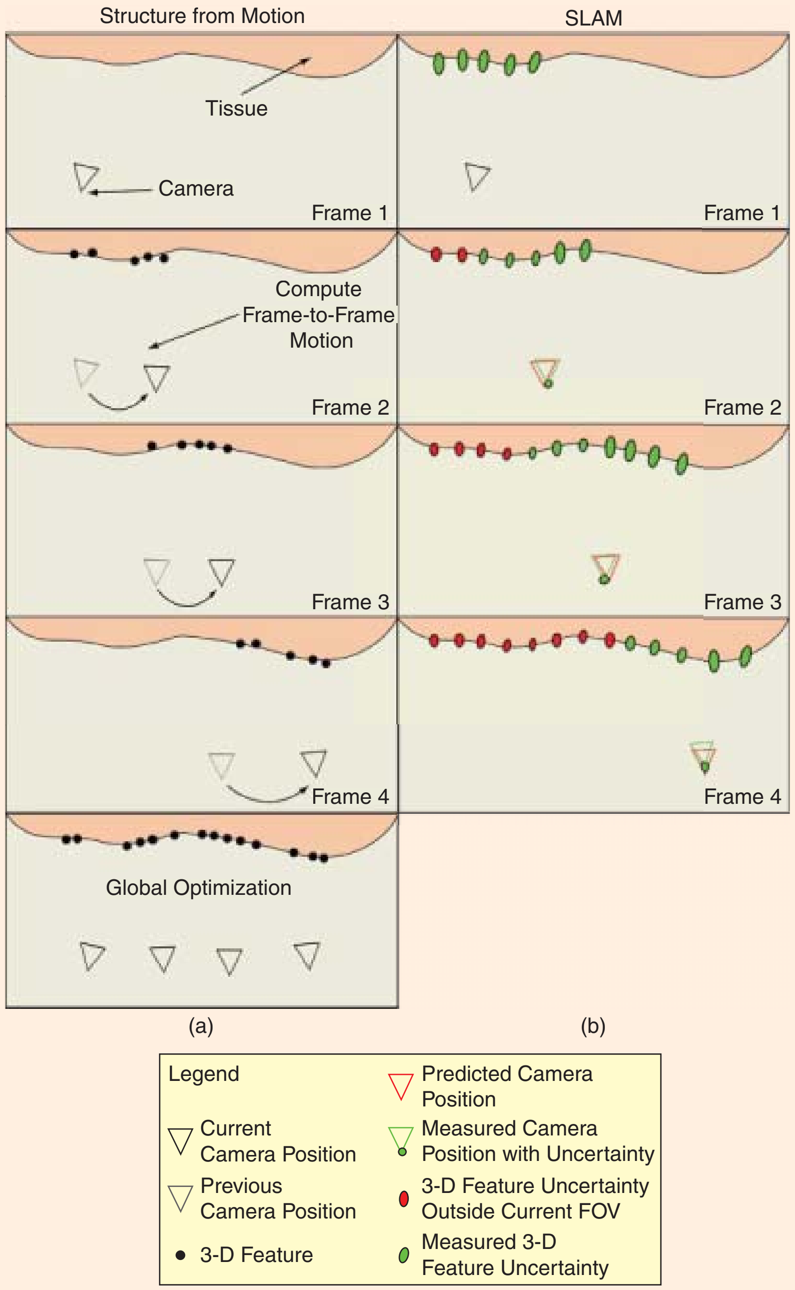
Image registration and frame-to-frame motion estimation can be performed in the image space by using direct alignment [53]–[56] or in feature space using region matching [57]–[60]. Direct alignment uses every pixel in the image and is well suited to environments with sparse regions of interest. However, it requires a large image overlap, suffers from the aperture problem, and can be affected by specular highlights. Operating in feature space enables registration with smaller image overlap and nonsequential matching. Camera motion is estimated by minimizing the equation [4] defined by the motion model.

The motion model describes the assumptions made about the structure and geometry of the environment. It defines the mathematical relationship between pixels in images captured from different locations. In MIS, planar models have been used on a variety of organs [59], [61]–[64] (see Table 2), cylindrical models for the esophagus and colon [53]–[56], and full projective models for the abdomen, colon [57], heart [58], and bladder [60]. The main problem with structure from motion is error propagation caused by the frame-to-frame camera motion estimation. Small errors propagate over time and can cause inaccuracies in the camera and structure estimations. This problem can be addressed using global optimization.

Global optimization is the use of batch operations or bundle adjustment to register multiple images together and find the optimal set of transformations that minimizes error and removes drift. Global optimization with multiple images can be computationally expensive, making it inappropriate for online in vivo, in situ applications. Nevertheless, it is suited to offline applications [56], [59], [61].

Scene reconstruction is the process of generating a model of the tissue structure. Given the estimated positions of the camera, scene reconstruction can be performed by matching regions of interest between images. The matched regions are triangulated to estimate 3-D points relative to the camera. These points can be meshed or interpolated to create a model of the tissue structure.

The work described above is based on the assumption that the MIS environment is static. Nonrigid structure from motion has been proposed for tracking faces [65] and clothing [66]. These techniques are based on the factorization method and shape basis representation. They are not suitable for real-time applications as the deformation is dealt with in an offline, global optimization step. Nonrigid structure from motion has been applied to the heart [67]. However, it is used to deal with residual motion when constructing a static cardiac surface at a preselected point in the cardiac cycle and not to generate a deforming surface model.



[FIG4] Illustration of structure and camera motion estimation. (a) Structure from motion with frame-to-frame estimation and global optimization. (b) SLAM with sequential incremental long-term mapping, uncertainty estimates, motion prediction, and state updates.

SLAM

SLAM has its origin in autonomous robotic navigation. It is designed to solve the problems of consistent incremental environment mapping and localization of a robot within the map. Previously, these had been treated as separate problems where either the map or robot location is assumed to be known. This approach was unsuccessful as neither can be known for certain due to noise in sensor measurements. The solution is to formulate mapping and localization into a single state estimation problem within a probabilistic framework. Originally developed for laser ranger finders and sonar, SLAM has been reformulated for cameras [68].

In MIS, SLAM has been applied to the abdomen [69]–[71], esophagus [72], and sinus [73] (in conjunction with preoperative data) where deformation and tissue motion is minimal.

Figure 5 shows the results of SLAM when applied to laparoscopic surgery, illustrating the 3-D map and camera position. The fulcrum effect of the laparoscope is clearly visible. In MIS, the goal is to localize the laparoscope camera and build a map of the tissue surface. A typical feature-based SLAM system is illustrated in Figure 4. The SLAM system alternates between a prediction step, where the motion of the camera is blindly predicted, and an update step, where the map is measured relative to the camera. A vision SLAM system consists of a state vector, a probabilistic framework, feature initialization, a prediction model, and a measurement model.

The state vector contains a map and the position of the laparoscope camera. The map contains the 3-D *xyz* position of a set of features or points in the environment. The camera's position is represented by the *xyz* position and roll, pitch, yaw rotations. In addition, this state vector contains the velocity and angular velocity of the camera. Real-time performance has been demonstrated [68] on sparse maps containing 100 features with full covariance.

The probabilistic framework in SLAM enables uncertainty or noise in the system to be modeled. The framework represents the joint probability between the position of the camera and the features in the map at a given point in time. It therefore corresponds to the current estimate of the state vector and the uncertainty in the state estimation. In MIS, the extended Kalman filter (EKF), which assumes Gaussian noise, has been employed [69]–[72], [74]. The uncertainty in the state estimate is represented in a covariance matrix, which describes the variance from the estimate. In the wider SLAM community, a variety of probabilistic frameworks have been implemented including unscented Kalman filters and Rao-Blackwellized particle filters (FastSLAM) [75].

Features initialization is dependent on the optical setup. In stereoscopic systems [69], [71], [74], features are matched in the left and right images and the 3-D position is triangulated relative to the camera. In monocular systems, the 3-D position is estimated by matching features temporally and requires the camera motion to be estimated. This is estimated using inverse depth [70], [72] or structure from motion [73]. SLAM uses a full covariance matrix between all features in the map to enable map convergence. For real-time performance, the size of the map is restricted and feature initialization should be carefully managed.

The prediction model or motion model describes how the camera is expected to move. This model contains the following two elements:

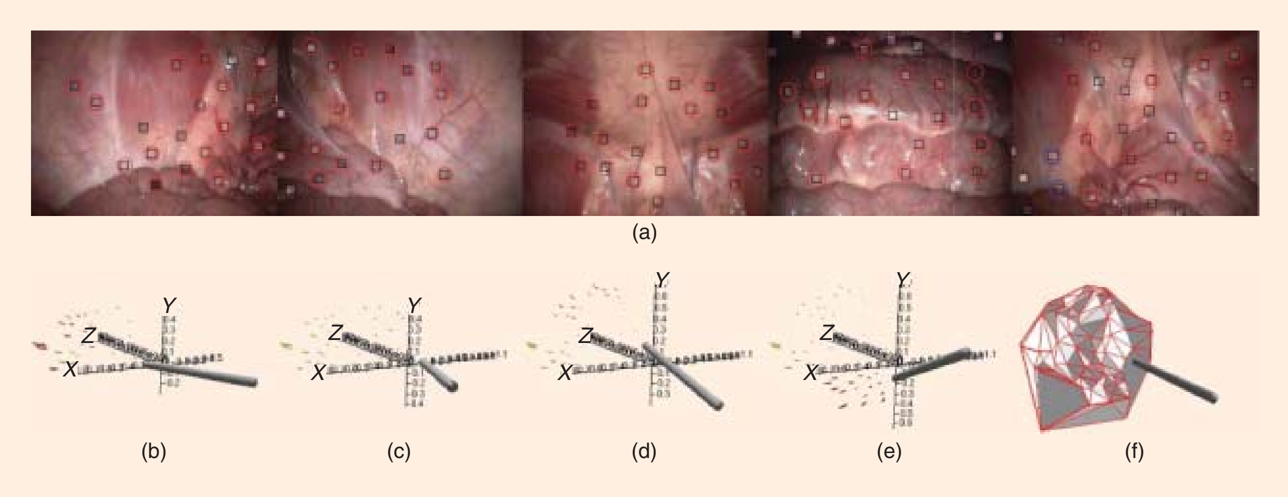
1) The deterministic element is where the motion is estimated based on a sensor (e.g., odometry) or an assumption. In [69], [70], and [72], a constant velocity constant acceleration model is assumed.

2) The stochastic element, which is a probability distribution represented by a Gaussian or collection of particles. It represents the unknown motion that cannot be easily modeled.

A constant velocity, constant acceleration motion model assumes the camera motion will be smooth. This assumption can be violated in both handheld MIS and robotic-assisted MIS, thus leading to system failure. The motion of a rigid laparoscope is limited by the fulcrum effect that may help to constrain the problem.

In the update step, the predicted state is compared to the measured state. The measurement model provides a means of measuring the current state of the system. SLAM measures the location of features in the map relative to the camera. In stereo SLAM, visible features are compared in 3-D by stereo region matching and triangulation, while in monocular SLAM visible features are projected onto the camera image plane and regions are matched using measurements in the 2-D image plane.

SLAM is a recent success story in mobile robotics that is also establishing its role for image-guided surgery, largely due to its probabilistic foundations and real-time capabilities. Unlike structure from motion, it is naturally suited to returning to previously visited areas and does not require a batch process to converge to an accurate estimation of the environment structure. Practical future work in the application of SLAM to MIS



[FIG5] Laparoscopic SLAM as applied to the abdominal MIS. (a) Laparoscopic video with tracked regions (squares) and projected uncertainly (circles). Laparoscope position and (b)–(e) 3-D sparse map of tissue with position uncertainties and (f) 3-D surface mesh of tissue.

will be focused on creating denser maps covering larger areas, identifying more robust long-term features, developing motion models better suited to rapid motion, and recovering from failure. However, the main challenge in the application of SLAM to MIS is the theoretical treatment of deformation.

SLAM has been widely applied to nonstatic civil environments where motion is caused by people and cars. Nonstatic motions are treated as outliers. Outliers can be identified using approaches such as RANSAC [44]. This assumes a global rigidity model and identifies outliers as features that do not fit to the model. This approach relies on parts of the environment being static that may not be the case in MIS. In [77], however, moving objects (cars) are identified and incorporated into the probabilistic framework of SLAM. This work demonstrates that SLAM can be applied without the full static assumption by explicitly creating motion models for moving objects. As we have seen in the section "Tissue Morphology Modeling," it is possible to estimate motion models representing the morphology of deforming tissue. Future work on deforming SLAM will investigate the incorporation of morphological models into the probabilistic framework.

The output from SLAM is generally a sparse set of 3-D points representing the structure of the environment. These points can be meshed to create a solid model shown in Figure 5(f). Textures from the laparoscopic video can be applied to make the model visually accurate.

MONOCULAR AND STEREO SYSTEMS

Structure from motion and SLAM can be used with either monocular or stereo cameras. Monocular systems are commonly used in operating theatre. However, the number of stereoscopic systems is steadily increasing particularly for robotic-assisted MIS. Ideally, the integration of computer vision into the surgical theatre will operate with existing monocular laparoscopes, however, the significant drawback of monocular vision is that acquiring depth information requires camera motion or fiducials of a known size. Therefore, the application of monocular vision in MIS is more limited than stereo.

VALIDATION AND VERIFICATION

Validation is a crucial step in the evaluation of the discussed methods. Practically, the validation process is challenging due to a lack of ground truth for in vivo cases. Experiments are usually performed on numerically simulated data or on phantom models. The ideal metric for measuring error should be Euclidian distances in metric 3-D space or in the projected 2-D image plane. However, for algorithms where rotations need to be evaluated, as with mosaicing, the exact method for measurement is less well defined [59]. Qualitative evaluations using physiological signal frequency comparisons have been used in the literature [14], [36].

Computer simulations are used to test the numerical stability of algorithms under different levels of modeled noise to establish the baseline performance [41], [58], [60], [69]. However, simulations are not capable of modeling all noise sources and the complexity of the MIS setup. Therefore, more realistic phantom experiments with known ground truth geometry and motion characteristics are used [41], [58], [60], [74], [76].

In practice, the ground truth for phantom models can be obtained using tomographic scanning and reconstruction techniques or surface scanning using range finders [58], [60], [73], [74], [76]. A practical challenge is to ensure the structural integrity of the model during ground truth acquisition. This is particularly difficult for dynamic models, where the model morphology must be consistently repeatable and synchronized between modalities [28]. Repeatable dynamic motions can be achieved by a combination of mechanical devices and signal generators [19], [28], [41]. High contrast fiducial markers are typically embedded in the phantom enabling registration between the experimental and ground truth coordinate systems. The quality of the resulting alignment is of crucial importance to the values obtained during validation and controlling the error in the ground truth to measurement registration is an important consideration.

Ground truth for the camera or surgical tools can be obtained using optical trackers or electromagnetic tracking devices [73], [74]. They require hand-eye calibration to relate the tracking device and the camera coordinate systems. In addition, controlling the error propagation between the optical, camera, and phantom model coordinate systems can often be a practical challenge that needs to be handled with care.

For better visual fidelity, a cadaver can be used in experiments, however, the ground truth for this is difficult to obtain and maintain due to gradual changes in tissue property [19], [74], [76]. The same problems arise during in vivo and wet lab experiments with animal studies. In these cases, structural and morphological ground truth is not available and results are usually presented to qualitatively demonstrate practical feasibility rather than metric measurements. Some experimental analysis may be performed by obtaining user feedback [34], [35, [39] and by comparisons with other physiological sensing equipment such as ECG signals [13], [18], [32], [35], [67].

Currently, there is no quality data repository providing a series of data sets for algorithm benchmarking, evaluation and comparison. This makes it particularly difficult for research centers without established MIS infrastructure to work in this area. To address this problem, we have introduced a database containing video data, calibration information, and ground truth data http://vip.doc.ic.ac.uk/vision.

TECHNICAL CHALLENGES AND CLINICAL APPLICATIONS

The future of navigation and control in robotic-assisted MIS is in the intelligent use of preoperative and intraoperative patient specific data. For intraoperative guidance and applying image guided, dynamic active constraints to avoid critical anatomical structures, it is necessary to develop fast and accurate techniques for 3-D surface reconstruction and motion estimation in situ. However, the development of computer vision techniques for these dynamic and nonrigid surgical scenes remains challenging.

The robustness of computer vision in MIS is affected by a number of factors including the paucity of features, specular highlights, rapid camera motion, small baseline, tissue deformation, surgical smoke, and occlusion. One of the major challenges is the theoretical treatment of tissue deformation, in particular, when combined with camera motion. New methods are required to adapt to the changing environment and to understand the dynamics of the structural morphology to anticipate risks and apply motion prediction.

Tissue motion caused by the respiratory and cardiac cycles can be modeled using periodic and quasi-periodic models. This is particularly important for highly dynamic procedures around the beating heart where motions arising from the cardiac and respiratory cycles affect the stability of the operating field. In these cases, an important control issue to consider is motion compensation, where the robotic tools are synchronized with the physiological motion to cancel out its rhythmic components. In cardiothoracic surgery, despite the use of mechanical stabilizers the anastomosis site can be unstable and motion compensation is required facilitate less invasive procedures such as TECAB [36], [50]. For the effective deployment of motion compensation, the operating frequency of the robotic device control must be determined to avoid redundancy and signal aliasing. Some preliminary studies indicate that this may be in the region of 100 Hz, which requires fast intraoperative processing. In fact the frequency of operation required by the computer-integrated surgical system to update information or robotic control needs to be identified and accuracy requirements clearly defined for different applications [78].

Nonperiodic tissue deformation is likely to require the fusion of optical information with prior biomechanical or statistical anatomical models and patient specific information. The problem is complicated further by tissue-tool interaction and topological changes of the tissue due to dissection. There is a critical need for a synergy between the robotic instruments' interactions with tissue and the surgeon. For systems directed at orthopaedic surgery, for example, this can be achieved by imposing active constraints on the tool's motion by using the preoperatively acquired, segmented, and modeled patient data [79].

For soft-tissue procedures, the problem is significantly more complex, largely due to the deformation and dynamics of the anatomy during surgery. To impose control constraints on the robotic instruments and to establish "no go" zones for protecting delicate parts of the anatomy, patient specific data must be updated in vivo to reflect the current location and changes in anatomical structure. This requires 3-D surface recovery in real time and the subsequent augmentation of geometric and biomechanical models that are physically accurate. By incorporating biomechanical tissue properties, it may be possible, to accurately delineate critical anatomical structures and deliver tactile sensing to reflect the dynamic active constraints imposed. However, a major challenge of physical-based modeling such as finite element modeling is how to obtain the model parameters using information from medical images to conform to the appearance and behavior of real tissue. By considering the tissue deformation in real time, the model parameters may be updated to improve the most up-to-date anatomical representation. The modeled tissue can then be used for intraoperative simulations, establishing dynamic active constraints, and delivering tactile feedback through the surgical console.

Information regarding the computer-integrated system must be effectively presented to the surgeon with considerations for error and uncertainty in the data visualization. In image-guided surgery, AR is the most common form of data fusion. The clinical benefit of image guidance has been well recognized in neuro and orthopaedic surgeries where the operating field is stable and undergoes only limited deformation [80].

The main problem with implementing AR for surgical navigation in robotically assisted MIS is in the accurate alignment of the computer-generated images with the real world. Accurate alignment of the real and virtual objects depends on the accurate tracking of the position and orientation of the viewing source with respect to the anatomy of interest. The complexity of tissue deformation during surgery imposes significant challenges to the AR display and it is a major factor that limits the more widespread use of AR for surgical guidance in soft-tissue procedures. In particular, deformation inhibits two important aspects of navigation: 1) recovery of the motion and the location of the imaging device with respect to the tissue and 2) the computation of the relationship between the preoperative model of the anatomy and its intraoperative status. The incorporation of 3-D shape recovery from stereo video sequences provides the possibility of AR being used for robotic-assisted laparoscopic surgeries. An important area of work is how to extend the current state of the art in localization techniques to handle deformable environments.

Human computer/robot interaction is another important part of future MIS platforms. Developing interfaces for the surgical theatre is challenging as the surgeons use their hands to perform surgery, making traditional interfaces such as keyboards and mice inappropriate. Foot peddles offer an additional source of input, however, they are limited in their range of input and in [81], it has been shown that voice control can be employed to position the endoscope. Eye-gaze tracking and brain-machine interfaces are elegant solution to the interface problem and have the potential to provide more information than traditional techniques, such as the focus and attention of the surgeon. This information have been exploited for visual servoing in [82] and for motion compensation in [83]. Developing intuitive interfaces for surgery can be challenging as surgical workflow can vary greatly between surgeons. It is envisaged that for complex image-guided procedures, a new profession of surgical analysts may be created in the future.

SUMMARY

Advanced surgical techniques such as image-guided navigation with intraoperative motion stabilization and dynamic active constraints have the potential to change the current functional capabilities of MIS. For these techniques to be successful in complex MIS procedures, accurate recovery of 3-D tissue structure and morphology, as well as camera motion estimation are important prerequisites. In this tutorial, we have outlined the current approaches to estimating this information using laparoscopic cameras. We have reviewed optical methods from camera models to tissue morphology recovery techniques for robotic guidance. This is an active research area that has witnessed a significant amount of research output in recent years. It is anticipated that with its maturity, the information derived will play a pivotal role in the future of image-guided or robotic-assisted MIS.

AUTHORS

Peter Mountney (peter.mountney05@imperial.ac.uk) received his M.Eng. degree (Hons) in computer science from the University of Bristol in 2003. He is currently a Ph.D. candidate at Imperial College. His research interests are in the use of laparoscopic cameras as a means of providing measurement of tissue structure during MIS. His current work focuses on image-guided surgery, soft tissue tracking, 3-D estimation, tissue modeling, and SLAM.

Danail Stoyanov (danail.stoyanov@imperial.ac.uk) received the Ph.D. degree from Imperial College London in 2006 for his work on soft-tissue 3-D structure and motion recovery in robotic-assisted MIS. He is currently a Royal Academy of Engineering/EPSRC Research Fellow at the Institute of Biomedical Engineering, Imperial College London. His main research interests are in computer vision, image-guided surgery, and surgical robotics.

Guang-Zhong Yang (g.z.yang@imperial.ac.uk) received the Ph.D. degree in computer science from Imperial College London. He is the founding director of the Royal Society/Wolfson Medical Image Computing Laboratory, research director of the Institute of Biomedical Engineering, and cofounder of the Hamlyn Centre for Robotic Surgery and Wolfson Surgical Technology Laboratory at Imperial College London. His main research interests include medical imaging, sensing, and robotics.

REFERENCES

[1] M. J. Mack, "Minimally invasive and robotic surgery," *J. Amer. Med. Assoc.*, vol. 285, no. 7, pp. 568–572, 2001.

[2] R. McKinlay, M. Shaw, and A. Park, "A technique for real-time digital measurements in laparoscopic surgery," *Surg. Endosc.*, vol. 18, no. 4, pp. 709–712, 2004.

[3] R. J. Damiano, "Robotics in cardiac surgery: The emperor's new clothes," *J. Thorac. Cardiovasc. Surg.*, vol. 134, no. 3, pp. 559–561, 2007.

[4] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[5] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 1106–1112.

[6] E. Prados and O. Faugeras, "Shape from shading: A well posed problem?," in *Proc. IEEE Computer Vision and Pattern Recognition*, San Diego, California, 2005, vol. 2, pp. 870–877.

[7] J.-Y. Bouguet. (2004). Camera calibration toolbox for Matlab [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[8] J. L. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman and Company, 1982.

[9] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.

[10] H. U. Rashid and P. Burger, "Differential algorithm for the determination of shape from shading using a point light source," *Image Vis. Comput.*, vol. 10, no. 2, pp. 119–127, 1992.

[11] F. Mourgues, F. Devernay, G. Malandain, and È. Coste-Manière, "3-D reconstruction of the operating field for image overlay in 3-D-endoscopic surgery," in *Proc. Int. Symp. Augmented Reality*, New York, 2001, pp. 191–192.

[12] G. Hager, B. Vagvolgyi, and D. Yuh, "Stereoscopic video overlay with deformable registration," in *Proc. Medicine Meets Virtual Reality*, 2007.

[13] M. Sauvée, A. Noce, P. Poignet, J. Triboulet, and E. Dombre, "Three-dimensional heart motion estimation using endoscopic monocular vision system: From artificial landmarks to texture analysis," *Biomed. Signal Process. Contr.*, vol. 2, no. 3, pp. 199–207, 2007.

[14] R. Ginhoux, J. A. Gangloff, M. d. Mathelin, L. Soler, M. M. A. Sanchez, and J. Marescaux, "Beating heart tracking in robotic surgery using 500 Hz visual servoing: Model predictive control and an adaptive observer," in *Proc. Int. Conf. Robotics and Automation*, 2004, pp. 274–279.

[15] C. H. Q. Forster and C. Tozzi, "Towards 3-D reconstruction of endoscope images using shape from shading," in *Proc. Brazilian Symp. Computer Graphics and Image Processing*, 2000, pp. 90–96.

[16] A. Tankus, N. Sochen, and Y. Yeshurun, "Perspective shape-from-shading by fast marching," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2004, pp. 43–49.

[17] D. Stoyanov, A. Darzi, and G.-Z. Yang, "Dense 3-D depth recovery for soft tissue deformation during robotically assisted laparoscopic surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2004, pp. 41–48.

[18] W. W. Lau, N. A. Ramey, J. Corso, N. V. Thakor, and G. D. Hager, "Stereo-based endoscopic tracking of cardiac surface deformation," in *Proc. Medical Image Computing and Computer Assisted Intervention*, St. Malo, France, 2004, vol. 3217, pp. 494–501.

[19] R. Richa, P. Poignet, and C. Liu, "Deformable motion tracking of the heart surface," in *Proc. Int. Conf. Intelligent Robots and Systems*, 2008, pp. 3997–4003.

[20] I. C. Albitar, P. Graebling, and C. Doignon, "Robust structured light coding for 3-D reconstruction," in *Proc. Int. Conf. Computer Vision*, 2007, pp. 1–6.

[21] T. Okatani and K. Deguchi, "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center," in *Proc. Computer Vision and Image Understanding*, 1997, vol. 66, pp. 119–131.

[22] H. Fuchs, M. A. Livingston, R. Raskar, D. Colucci, K. Keller, A. State, J. R. Crawford, P. Rademacher, S. H. Drake, and A. A. Meyer, "Augmented reality visualization for laparoscopic surgery," in *Proc. Medical Image Computing and Computer-Assisted Intervention*, 1998, vol. 1496, pp. 934–943.

[23] C. Wu, S. G. Narasimhan, and B. Jaramaz, "A multi-image shape-from-shading framework for near-lighting perspective endoscopes," *Int. J. Comput. Vis.*, vol. 86, no. 2–3, pp. 211–228, 2009.

[24] B. P. L. Lo, M. Visentini-Scarzanella, D. Stoyanov, and G.-Z. Yang, "Belief propagation for depth cue fusion in minimally invasive surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2008, pp. 104–112.

[25] H. Haneishi, T. Ogura, and Y. Miyake, "Profilometry of a gastrointestinal surface by an endoscope with laser beam projection," *Opt. Lett.*, vol. 19, no. 9, pp. 601–603, 1994.

[26] M. Hayashibe, N. Suzuki, and Y. Nakamura, "Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management," *Med. Image Anal.*, vol. 10, no. 4, pp. 509–519, 2006.

[27] D. Stoyanov, D. Elson, and G.-Z. Yang, "Illumination position estimation for 3-D soft-tissue reconstruction in robotic minimally invasive surgery," in *Proc. Intelligent Robots and Systems*, 2009, pp. 2628–2633.

[28] M. Visentini-Scarzanella, G. P. Mylonas, D. Stoyanov, and G.-Z. Yang, "I-Brush: A gaze-contingent virtual paintbrush for dense 3-D reconstruction in robotic assisted surgery," in *Proc. Medical Image Computing and Computer-Assisted Intervention*, 2009, pp. 353–360.

[29] D. Stoyanov, A. Darzi, and G.-Z. Yang, "A practical approach towards accurate dense 3-D depth recovery for robotic laparoscopic surgery," *Comput. Aid. Surg.*, 2005, vol. 10, pp. 199–208.

[30] T. P. Koninckx and L. Van-Gool, "Real-time range acquisition by adaptive structured light," *IEEE Trans. Pattern Anal. Machine Intell.*, 2006, vol. 28, pp. 432–445.

[31] J. Penne, K. Höller, M. Stürmer, T. Schrauder, A. Schneider, R. Engelbrecht, H. Feußner, B. Schmauss, and J. Hornegger, "Time-of-flight 3-D endoscopy," in *Proc. Medical Image Computing and Computer Assisted Interventions*, 2009, pp. 467–474.

[32] M. Gröger, T. Ortmaier, W. Sepp, and G. Hirzinger, "Tracking local motion on the beating heart," in *Proc. SPIE Medical Imaging Conf.*, San Diego, California, 2002, vol. 4681, pp. 233–241

[33] D. Stoyanov, G. P. Mylonas, F. Deligianni, A. Darzi, and G.-Z. Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2005, vol. 3750, pp. 139–146.

[34] P. Mountney, B. P. L. Lo, S. Thiemjarus, D. Stoyanov, and G.-Z. Yang, "A probabilistic framework for tracking deformable soft tissue in minimally invasive

surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2007, vol. 2, pp. 34–41.

[35] P. Mountney and G.-Z. Yang, "Soft tissue tracking for minimally invasive surgery: Learning local deformation online," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2008, pp. 364–372.

[36] T. Ortmaier, M. Gröger, D. H. Boehm, V. Falk, and G. Hirzinger, "Motion estimation in beating heart surgery," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 10, pp. 1729–1740, 2005.

[37] C. Wengert, L. Bossard, A. Häberling, C. Baur, G. Székely, and P. C. Cattin, "Endoscopic navigation for minimally invasive suturing," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2007, vol. 792, pp. 620–627.

[38] S. Giannarou, M. Visentini-Scarzanella, and G.-Z. Yang, "Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery," in *Proc. Int. Symp. Biomedical Imaging*, 2009, pp. 1059–1062.

[39] N. Masson, F. Nageotte, P. Zanne, M. d. Mathelin, and J. Marescaux, "Comparison of visual tracking algorithms on in vivo sequences for robot-assisted flexible endoscopic surgery," in *Proc. Engineering in Medicine and Biology Conf.*, 2009, pp. 5571–5576.

[40] A. Noce, J. Triboulet, P. Poignet, and E. Dombre, "Texture features selection for visual servoing of the beating heart," in *Proc. BioRob*, 2006, pp. 335–340.

[41] L. Ott, P. Zanne, F. Nageotte, M. d. Mathelin, and J. Gangloff, "Physiological motion rejection in flexible endoscopy using visual servoing," in *Proc. Int. Conf. Robotics and Automation*, 2008, pp. 2928–2933.

[42] V. Lepetit and P. Fua, "Monocular model-based 3-D tracking of rigid objects: A survey," *Found. Trends Comput. Graph. Vis.*, vol. 1, pp. 1–89, 2005.

[43] M. Gröger and G. Hirzinger, "Optical flow to analyse stabilised images of the beating heart," in *Proc. Int. Conf. Computer Vision Theory and Applications*, 2006, pp. 237–244.

[44] P. H. S. Torr and D. W. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *Int. J. Comput. Vis.*, vol. 24, no. 3, pp. 271–300, 1997.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[46] H. Bay, T. Tuytelaars, and L. van-Gool, "Surf: Speeded up robust features," in *Proc. European Conf. Computer Vision*, 2006, vol. 3951, pp. 404–417.

[47] C. N. Riviere, J. Gangloff, and M. d. Mathelin, "Robotic compensation of biological motion to enhance surgical accuracy," *Proc. IEEE*, vol. 94, no. 9, pp. 1705–1716, 2006.

[48] C. Riviere, A. Thakral, I. I. Iordachita, G. Mitroi, and D. Stoianovici, "Predicting respiratory motion for active canceling during percutaneous needle insertion," in *Proc. Engineering in Medicine and Biology Conf.*, 2001, pp. 3477–3480.

[49] R. Richa, A. P. L. Bo, and P. Poignet, "Motion prediction for tracking the beating heart," in *Proc. Engineering in Medicine and Biology Conf.*, 2008, pp. 3261–3264.

[50] R. Ginhoux, J. Gangloff, M. d. Mathelin, L. Soler, M. M. A. Sanchez, and J. Marescaux, "Active filtering of physiological motion in robotized surgery using predictive control," *IEEE Trans. Robot.*, vol. 21, no. 1, pp. 67–79, 2005.

[51] L. Cuvillon, J. Gangloff, M. d. Mathelin, and A. Forgione, "Toward robotized beating heart Tecabg: Assessment of the heart dynamics using high-speed vision," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2005, vol. 2, pp. 551–558.

[52] S. Misra, K. T. Ramesh, and A. M. Okamura, "Modeling of tool-tissue interactions for computer-based surgical simulation: A literature review," *Presence: Teleoperators Virtual Environ.*, vol. 17, no. 5, pp. 463–491, 2008.

[53] J. Zhou, A. Das, F. Li, and B. Li, "Circular generalized cylinder fitting for 3-D reconstruction in endoscopic image based MRF," in *Proc. IEEE Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8.

[54] E. J. Seibel, R. E. Carroll, J. A. Dominitz, R. S. Johnston, C. D. Melville, C. M. Lee, S. M. Seitz, and M. B. Kimmey, "Tethered capsule endoscopy. A low-cost and high-performance alternative technology for the screening of esophageal cancer and Barrett's esophagus," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1032–1042, 2008.

[55] S. Seshamani, W. Lau, and G. Hager, "Real-time endoscopic mosaicking," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2006, vol. 1, pp. 355–363.

[56] R. E. Carroll and S. M. Seitz, "Rectified surface mosaics," in *Proc. Int. Conf. Computer Vision*, 2007, pp. 1–8.

[57] D. Koppel, C.-I. Chen, Y.-F. Wang, H. Lee, J. Gu, A. Poirson, and R. Wolters, "Toward automated model building from video in computer-assisted diagnoses in colonoscopy," in *Proc. SPIE*, 2007, vol. 6509, no. 2.

[58] M. Hu, G. P. Penney, P. J. Edwards, M. Figl, and D. J. Hawkes, "3-D reconstruction of internal organ surfaces for minimal invasive surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2007, vol. 1, pp. 68–77.

[59] S. Atasoy, D. P. Noonan, S. Benhimane, N. Navab, and G.-Z. Yang, "A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2008, vol. 1, pp. 850–857.

[60] C.-H. Wu, Y.-N. Sun, and C.-C. Chang, "Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 7, pp. 1199–1211, 2007.

[61] R. Miranda-Luna, C. Daul, W. C. P. M. Blondel, Y. Hernandez-Mier, D. Wolf, and F. Guillemin, "Mosaicing of bladder endoscopic image sequences: distortion calibration and registration algorithm," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 2, pp. 541–553, 2008.

[62] S. Olijnyk, Y. H. Mier, W. C. P. M. Blonde , C. Daul, D. Wolf, and G. Bourg-Heckly, "Combination of panoramic and fluorescence endoscopic images to obtain tumor spatial distribution information useful for bladder cancer detection," *Progr. Biomed. Opt. Imag.*, vol. 8, no. 44, 2007.

[63] A. Behrens, "Creating panoramic images for bladder fluorescence endoscopy," *Acta Polytech. J. Adv. Eng.*, vol. 48, no. 3, pp. 50–54, 2008.

[64] T. Igarashi, H. Suzuki, and Y. Naya, "Computer based endoscopic image processing technology for endourology and laparoscopic surgery," *Int. J. Urol.*, vol. 16, no. 6, pp. 533–543, 2009.

[65] M. Brand, "Morphable 3-D models from video," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2001, vol. 2, pp. 456–463.

[66] L. Torresani, D. B. Yang, E. J. Alexander, and C. Bregler, "Tracking and modeling non-rigid objects with rank constraints," in *Proc. IEEE Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 493–500.

[67] M. Hu, G. P. Penney, D. Rueckert, P. J. Edwards, R. Bello, R. Casula, M. Figl, and D. J. Hawkes, "Non-rigid reconstruction of the beating heart surface for minimally invasive cardiac surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2009, pp. 34–42.

[68] A. J. Davison, I. Reid, N. Molton, and O. Stasse, "Monoslam: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Machine Intell.*, 2007, vol. 29, pp. 1052–1067.

[69] P. Mountney, D. Stoyanov, A. J. Davison, and G.-Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2006, vol. 1, pp. 347–354.

[70] O. Garcıa, J. Civera, A. Gueme, V. Munoz, and J. M. M. Montiel, "Real-time 3-D modeling from endoscope image sequences," in *Proc. Int. Conf. Robotics and Automation Workshop on Advanced Sensing and Sensor Integration in Medical Robotics*, 2009.

[71] P. Mountney and G.-Z. Yang, "Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping," in *Proc. Engineering in Medicine and Biology Conf.*, 2009, pp. 1184–1187.

[72] V. Castaneda, S. Atasoy, D. Mateus, N. Navab, and A. Meining, "Reconstructing the esophagus surface from endoscopic image sequences," in *Proc. Russian Bavarian Conf. Bio-Medical Engineering*, 2009.

[73] D. Burschka, M. Li, M. Ishii, R. Taylor, and G. D. Hager, "Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery," *Med. Image Anal.*, vol. 9, no. 5, pp. 413–426, 2005.

[74] D. Noonan, P. Mountney, D. Elson, A. Darzi, and G.-Z. Yang, "A stereoscopic fibroscope for camera motion and 3-D depth recovery during minimally invasive surgery," in *Proc. Int. Conf. Robotics and Automation*, 2009, pp. 4463–4468.

[75] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. Cambridge, MA: MIT Press, 2005.

[76] R. Richa, P. Poignet, and C. Liu, "Efficient 3-D tracking for motion compensation in beating heart surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2008, pp. 684–691.

[77] C. C. Wang, C. Thorpe, and S. Thrun, "Online simultaneous localization and mapping with detection and tracking of moving objects: Theory and results from a ground vehicle in crowded urban areas," in *Proc. Int. Conf. Robotics and Automation*, 2003, vol. 1, pp. 842–849.

[78] R. H. Taylor and D. Stoianovici, "Medical robotics in computer-integrated surgery," *IEEE Trans. Robot. Automat.*, vol. 19, no. 5, pp. 765–781, 2003.

[79] B. Davies, M. Jakopec, S. J. Harris, Y. Rodriguez, F. Baena, A. Barrett, A. Evangelidis, P. Gomes, J. Henckel, and J. Cobb, "Active-constraint robotics for surgery," *Proc. IEEE*, vol. 94, pp. 1696–1704, 2006.

[80] T. M. Peters, "Image-guidance for surgical procedures," *Phys. Med. Biol.*, vol. 51, no. 14, pp. R505–R540, 2006.

[81] L. K. Jacobs, V. Shayani, and J. M. Sackier, "Determination of the learning curve of the Aesop robot," *Surg. Endosc.*, vol. 11, no. 1, pp. 54–55, 1997.

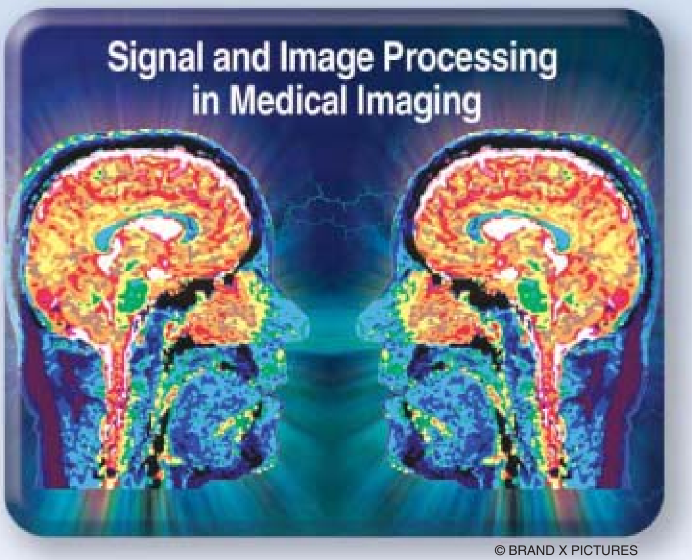
[82] D. P. Noonan, G. P. Mylonas, A. Darzi, and G.-Z. Yang, "Gaze contingent articulated robot control for robot assisted minimally invasive surgery," in *Proc. Intelligent Robots and Systems*, 2008, pp. 1186–1191.

[83] G. P. Mylonas, D. Stoyanov, F. Deligianni, A. Darzi, and G.-Z. Yang, "Gaze-contingent soft tissue deformation tracking for minimally invasive robotic surgery," in *Proc. Medical Image Computing and Computer Assisted Intervention*, 2005, vol. 3749, pp. 843–850.

[Miles N. Wernick, Yongyi Yang, Jovan G. Brankov, Grigori Yourganov, and Stephen C. Strother]

Machine Learning in Medical Imaging

[Drawing conclusions from medical images]



© BRAND X PICTURES

Statistical methods of automated decision making and modeling have been invented (and reinvented) in numerous fields for more than a century. Important problems in this arena include pattern classification, regression, control, system identification, and prediction. In recent years, these ideas have come to be recognized as examples of a unified concept known as machine learning, which is concerned with 1) the development of algorithms that quantify relationships within existing data and 2) the use of these identified patterns to make predictions based on new data. Optical character recognition, in which printed characters are identified automatically based on previous examples, is a classic engineering example of machine learning. But this article will discuss very different ways of using machine learning that may be less familiar, and we will demonstrate through examples the role of these concepts in medical imaging.

Machine learning has seen an explosion of interest in modern computing settings such as business intelligence, detection of e-mail spam, and fraud and credit scoring. The medical imaging field has been slower to adopt modern machine-learning techniques to the degree seen in other fields. However, as computer power has grown, so has interest in employing advanced algorithms to facilitate our use of medical images and to enhance the information we can gain from them.

Although the term *machine learning* is relatively recent, the ideas of machine learning have been applied to medical imaging for decades, perhaps most notably in the areas of computer-aided diagnosis (CAD) and functional brain mapping. We will not attempt in this brief article to survey the rich literature of this field. Instead our goals will be 1) to acquaint the reader with some modern techniques that are now staples of the machine-learning field and 2) to illustrate how these techniques can be employed in various ways in medical imaging using the following examples from our own research:

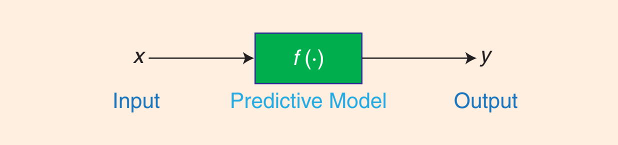
- CAD
- content-based image retrieval (CBIR)
- automated assessment of image quality
- brain mapping.

INTRODUCTION TO MACHINE LEARNING

In this brief tutorial, we will attempt to introduce a few basic techniques that are widely applicable and then show how these can be used in various medical imaging settings using examples from our past work in this field. For further

Digital Object Identifier 10.1109/MSP.2010.936730

1053-5888/10/$26.00©2010IEEE



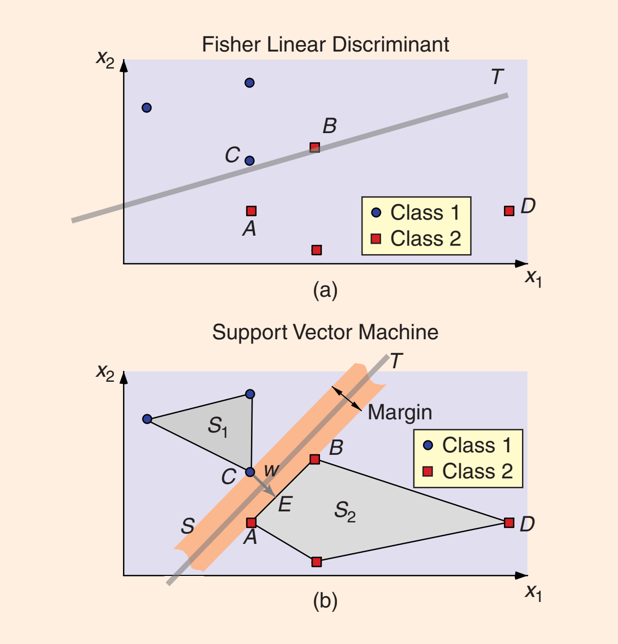
[FIG1] In supervised learning the predictive model represents the assumed relationship between input variables in x and output variable *y*.

information, interested readers should consult well-known introductions to machine learning, such as the excellent treatments in [1] and [2].

SUPERVISED LEARNING

In machine learning, one often seeks to predict an output variable y based on a vector $\mathbf{x}$ of input variables. To accomplish this, it is assumed that the input and output approximately obey a functional relationship $y = f(\mathbf{x})$, called the predictive model, as shown in Figure 1. In supervised learning, the predictive model is discovered with the benefit of training data consisting of examples for which both $\mathbf{x}$ and y are known. We will denote these available pairs of examples as $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, and we will assume that $\mathbf{x}$ is composed of n variables (called features), so that $\mathbf{x}_i \in \mathbb{R}^n$. In general, the output of the predictive model can be a vector (e.g., in multiclass classifiers), but for simplicity we will confine our attention to the case of scalar outputs.

Historically, a somewhat artificial distinction has sometimes been made between two learning problems: classification and regression. Classification refers to decision among a typically small and discrete set of choices (such as identifying a tumor as malignant or benign), whereas regression refers to estimation of a possibly continuous-valued output variable (such as a diagnostic assessment of disease severity y). If the choices in a classification problem are indicated by discrete numerical values (e.g., $y = +1$ for the class malignant and $y = -1$ for benign), then it is easy to see that classification and regression are represented equivalently by the model in Figure 1.



[FIG2] Fisher linear discriminant (LD) and the SVM. In this example, (a) the Fisher LD fails to separate two classes because training example *D* adversely influences decision boundary *T*. (b) The SVM defines the decision boundary using only points *A*, *B*, and *C*, called support vectors, and is not influenced at all by point *D*.

THE SUPPORT VECTOR MACHINE CLASSIFIER: A MAXIMUM-MARGIN APPROACH

Let us consider the simple pattern classification problem depicted in Figure 2, in which the goal is to segregate vectors $\mathbf{x} = (x_1, x_2)^T$ into two classes by using a decision boundary T. Let us employ a linear model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, so that T is a line in this two-dimensional example. Traditionally, the model's parameters ($\mathbf{w}$ and b in this case) have been determined using classical criteria such as least squares or maximum likelihood. Figure 2 illustrates how such an approach (in this case, a Fisher discriminant) can easily fail, particularly when the method's distributional assumptions are violated. In Figure 2(a), data point D adversely influences the Fisher discriminant boundary, causing misclassification of point B even though point D lies very far from Class 1, and perhaps should not be granted this degree of influence.

The support vector machine (SVM) [2], discovered by Vapnik, resolves this shortcoming by defining the discriminant boundary only in terms of those training examples that lie dangerously close to the class to which they do not belong. This idea is understood most easily in a situation such as the one shown in Figure 2, in which the two classes are strictly separable by a linear decision boundary, as explored by Wernick in [3]. In this case, a separating line that maximizes the margin between the two classes can always be found as follows:

1) Draw the convex hull of each class of data points (imagine stretching a rubber band around each group of points; call these regions S_1 and S_2).

2) Find the points C and E at which regions S_1 and S_2 have their closest approach.

3) Draw the perpendicular bisector of the line segment connecting points C and E to obtain the decision boundary T.

Step 2 is accomplished by solving a quadratic programming (constrained optimization) problem using standard approaches [3]. In linear classifiers, vector $\mathbf{w}$ is called the discriminant vector.

In the terminology of the SVM, points A, B, and C in Figure 2 are called support vectors, a term derived from an analogy to mechanics. If points A, B, and C in Figure 2 were physical supports, they would be sufficient to provide mechanical stability to slab S sandwiched between them.

It is evident that the support vectors are the only examples from the training data that explicitly define the model. Specifically, for a particular test example $\mathbf{x}$, one can write the model in terms of the support vectors as follows:

$$f(\mathbf{x}) = \sum\nolimits_{i \in I_s} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b, \tag{1}$$

in which the summation includes only the training examples $\mathbf{x}_i$ that are support vectors, and α_i are coefficients determined as Lagrange multipliers in the optimization procedure.

The benefits of the SVM approach are that the classifier concentrates automatically on examples that are difficult to classify (points A, B, and C); and the calculation in (1) scales with the number of support vectors rather than the dimension of the space (which in some problems is very large). In addition, SVM can be shown to balance training error and model complexity, thereby avoiding overfitting, a pitfall in which the model is too finely tuned to the training examples and fails to perform well on new data. This approach is called structural risk minimization [4].

The formulation described thus far does not allow for the possibility that the two classes cannot be entirely separated by a linear boundary. However, this situation is readily addressed by introducing slack variables into the quadratic optimization problem, thus allowing a minimal number of the training data to be misclassified. In addition, SVM can be easily adapted to accomplish regression instead of classification by using a so-called ε-insensitive cost function [2].

NONLINEAR MODELS: THE KERNEL TRICK

An important breakthrough in machine learning has been the recognition of the so-called kernel trick [2], which provides a simple and broadly applicable means to obtain a nonlinear model from any linear model based on inner products. Even classical techniques, such as the Fisher discriminant or principal component analysis, can be turned easily into flexible nonlinear techniques via the kernel trick.

To understand the kernel trick, consider the following hypothetical series of steps as applied to turn the linear SVM into a nonlinear technique. Suppose we were to first apply a nonlinear transformation Φ to each input vector $\mathbf{x}_i$ from the training set and then train a linear classifier to distinguish these classes of transformed vectors $\Phi(\mathbf{x}_i)$. Separability will be enhanced if the dimension of the transform space is higher than that of the original space, and indeed the transformation's dimension need not be finite.

At first glance, transforming each input vector into a space of high dimension might appear impractical. However, the kernel trick recognizes that the desired result can be obtained without actually performing the transformation. This can be seen by applying the transformation Φ and then applying the SVM model in (1). After transformation, (1) becomes

$$f(\mathbf{x}) = \sum\nolimits_{i \in I_s} \alpha_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + b. \tag{2}$$

Note that the transformation Φ appears in (2) only in the form of an inner product $K(\mathbf{x}_i, \mathbf{x}) \triangleq \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x})$, so that (2) can be rewritten as

$$f(\mathbf{x}) = \sum\nolimits_{i \in I_s} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{3}$$

Therefore, we can see that it is never actually necessary to compute Φ (or even to define it explicitly). Instead it is sufficient simply to define the kernel function $K(\cdot, \cdot)$, and it can be shown that any symmetric positive semidefinite function will suffice. Commonly used kernel functions in machine learning include radial basis functions (Gaussians) and polynomials. Intuitively, the effect of the kernel is to measure the "similarity" between a test vector $\mathbf{x}$ and each of the support vectors $\mathbf{x}_i$; these similarities are then used in to obtain the output result. Vectors belonging to one of the classes are presumably most "similar" to the support vectors belonging to that class, hence these similarity values convey the needed information. The key point to remember is that these similarity comparisons are made only in relation to the support vectors, which are difficult examples that lie near the discriminant boundary. We will see visual examples of these support vectors later in the setting of mammography.

RELEVANCE VECTOR MACHINES: BAYESIAN LEARNING AND SPARSITY CONSTRAINTS

An important successor of SVM is the so-called relevance vector machine (RVM), developed by Tipping [5]. We have found RVM to perform extremely well in several medical imaging applications, usually with much lower computational cost than alternative methods including SVM. The RVM emphasizes *s*parsity (i.e., reduced model complexity), and thus is closely related to ideas of compressed sensing [6]. Like SVM, RVM uses a subset of the training data called relevance vectors, but usually there are far fewer relevance vectors than support vectors.

Like SVM, RVM starts with a kernel model

$$f(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i), \tag{4}$$

however, whereas SVM is based on the maximum-margin principle, RVM instead takes a Bayesian approach. RVM assumes a Gaussian prior on the kernel weights w_i, which are assumed to have zero mean and variance a_i^{-1}. RVM further assumes a gamma hyperprior on a_i^{-1}. The net effect of these modeling choices is that the overall prior on the kernel weights w_i is a multivariate t-distribution. Because this distribution is tightly concentrated about the axes of the w_i space, the prior encourages most values of w_i to be nearly zero. Thus, in the end, the summation in involves only a few nonzero terms, and the associated training examples are called *relevance vectors*. By this mechanism, overfitting is generally avoided, and computation times for RVM are relatively low. Surprisingly, in spite of its advantages, RVM has been used relatively infrequently in medical imaging, particularly in comparison with the better-known SVM approach.

While RVM and SVM both base their decisions entirely on a subset of the training data (the relevance vectors in RVM; the support vectors in SVM), these subsets are usually quite different. Support vectors are always examples lying near the decision boundary, while relevance vectors are usually spread throughout the distribution. We will see this difference later in the context of mammography.

Unfortunately, RVM does not have a simple geometrical interpretation as SVM does, therefore we will not show a graphical example in this article; instead we refer the reader to [5], which contains several nice illustrations.

MACHINE LEARNING HAS SEEN AN EXPLOSION OF INTEREST IN MODERN COMPUTING SETTINGS SUCH AS BUSINESS INTELLIGENCE, DETECTION OF E-MAIL SPAM AND FRAUD, AND CREDIT SCORING.

STATISTICAL RESAMPLING FOR ROBUSTNESS AND EVALUATION

Statistical resampling [7] refers to a family of techniques that are used to evaluate performance and improve robustness of machine learning models and to estimate statistical significance levels. Although resampling receives less attention than predictive models, it is at least as important.

Machine learning differs from classical decision and estimation theory principally in its emphasis on problems where one's only knowledge of the data's underlying distributions comes from the data themselves. In this setting, statistical significance testing cannot be approached in the traditional way because the null distribution is unknown. Fortunately, an empirical estimate of the null distribution can be readily obtained by permutation resampling.

To understand permutation resampling, consider a situation in which there are two sets of data, ω_1 and ω_2, and we wish to test some hypothesis, such as that their means are identical. Since we do not know in truth whether ω_1 and ω_2 obey the same distribution (or even the form of their distributions), we cannot directly assess significance. However, we can create an empirical null distribution by permuting the labels on the data, i.e., deliberately creating two data sets in which the data from ω_1 and ω_2 are mixed. Note that it is often important that just the labels and not the data themselves be permuted (e.g., in time series problems). By permuting the data in every possible way (or at least in some reasonably large number of random ways), we can obtain example data in which we know that the two groups obey identical distributions, thus characterizing the null hypothesis.

Another central role played by resampling is in solving the following problem of model validation: If we train our model on all our available data, then there are no data left for testing the model or optimizing its parameters. The predominant resampling methods used in this regard, which both require independent, identically distributed (i.i.d.) resampling objects, are cross validation and bootstrap methods. In k-fold cross validation, the data set is divided randomly into k groups; $(k-1)$ of these groups are used to train the model, and one is reserved for testing. This process is performed k times (once for each held out group), then the results are combined, often by averaging. In the basic bootstrap, the data are instead trained on a set of N data examples obtained by sampling randomly with replacement from the entire data set of N. By chance, some examples will not be selected into the training set, and these are reserved for testing. As in cross validation, the process is repeated and the results combined by averaging.

The basic bootstrap is known to reduce the variance of estimated prediction accuracy at the expense of downward bias (i.e., the basic bootstrap provides pessimistic performance estimates). This is remedied by the .632 bootstrap, which utilizes a bias correction term, and the more modern .632+ bootstrap [8], which additionally attempts to account for bias due to overfitting. In problems where an empirical null distribution is obtained using permutations, the empirical distribution of the alternative hypothesis can often be obtained using the bootstrap.

Statistical resampling is widely used not only to test predictive models, but also to improve their performance. Examples of this are bootstrap aggregation (bagging) techniques and the nonparametric, prediction, activation, influence, reproducibility, resampling (NPAIRS) framework in neuroimaging [9], which is explained later in this article.

CAD FOR MAMMOGRAPHY

CAD has been an active research area for decades, so we will not attempt to provide a comprehensive survey of the literature. Interested readers should consult basic reviews of CAD for mammography, such as [10] and [11].

Perhaps CAD's greatest success is in breast imaging. Studies have shown that having two radiologists read the same mammogram can lead to significantly higher sensitivity in cancer screening, but at the expense of increased workload and cost. CAD software can serve as a surrogate "second reader," with the aim of improving radiologists' diagnostic accuracy at lower cost.

CAD encompasses computer-aided detection (CADe), in which the computer alerts the radiologist to potential lesions; and computer-aided diagnosis (CADx), in which the computer predicts the likelihood that a lesion is malignant.

CAD schemes typically consist of the following key steps: 1) apply automated image analysis to extract a vector of quantitative features to characterize the relevant image content and 2) apply a pattern classifier to determine the category to which the extracted feature vector may belong.

Automatically extracted image features can include image contrast, and features based on geometry, morphology, and texture. In addition, there may be other forms of available information about the patient. Machine-learning methods that have been employed range from linear discriminant (LD) analysis, fuzzy logic techniques, neural networks, and committee machines, to the more recent kernel-based methods (e.g., SVM and RVM) explained earlier in this article.

In the following, we describe two examples of machine learning for CAD in digital mammography drawn from our own research: detection (CADe) and classification (CADx) of clustered microcalicifications.

CADe: MICROCALCIFICATION DETECTION

Microcalcifications (MCs) are tiny deposits of calcium that appear as bright spots in mammograms (see Figure 3). Clustered MCs can be an important indicator of breast cancer, appearing in 30–50% of cases. Individual MCs are sometimes difficult to detect due to their variation in shape, orientation, brightness and size (typically, 0.05–1 mm), and because of the confounding texture of surrounding breast tissue. Microcalcification detection has been an intensive target of investigation (e.g., [12]). Modern machine-learning approaches have proven very effective in this application, as we explain next.

SVM Detector

In [13], we trained an SVM to decide at each location within a mammogram whether an MC was present ("MC present" class) or absent ("MC absent" class) based on a small region of interest (ROI) surrounding that point. The SVM was trained using "MC present" ROIs identified by expert radiologists (see Figure 4).

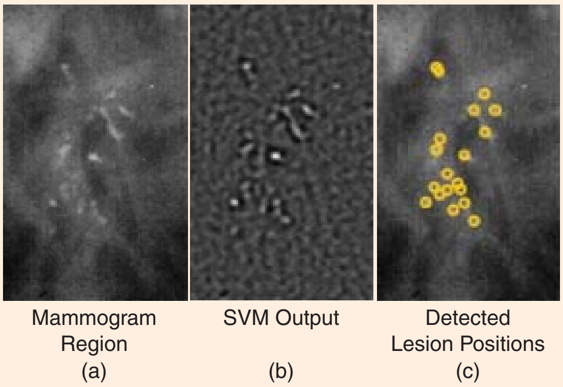
The MCs typically occupy only a small fraction of a mammogram, so there are more ROIs with "MC absent" than with "MC present." To take advantage of this, we developed a successive enhancement learning (SEL) procedure that improves the predictive power of the SVM classifier. In SEL, SVM training is adjusted iteratively by selecting the most representative "MC absent" examples from all the available training images while keeping the total number of training examples small.

Based on a set of test mammograms, we demonstrated the SEL-SVM method to achieve the best performance among several leading methods in the literature as measured by the free-response receiver operating characteristic (FROC) curve, a plot of detection probability versus the average number of false positives (FPs) per image (Figure 5). Figure 3 shows a portion of an example image and the corresponding SVM output.

RVM Detector

Computation time can be a critical issue in mammography, where the image can contain as many as 3,000 × 5,000 pixels that must be evaluated. While the SVM achieves outstanding detection performance, it can be very time consuming because the number of support vectors can be large. To address this issue, in [14] we developed an approach based on the RVM (explained earlier), which yields a very sparse decision function, leading to significant computational savings, while yielding similar detection performance to the SVM.

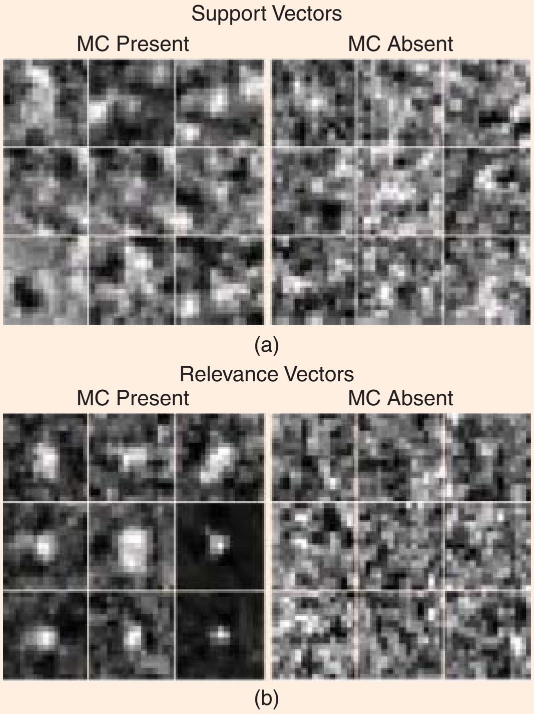
To further accelerate the algorithm, we explored a two-stage classification approach in which we used a computationally inexpensive linear RVM classifier as an initial triage step to quickly eliminate non-MC pixels, then a nonlinear RVM classifier to detect MCs among the remaining pixels. Our results demonstrated that the RVM approach achieved nearly identical detection accuracy to the SVM at 35 times less computational cost.



[FIG3] (a) Example mammogram containing microcalcifications. (b) Output *y* of SVM detector. (c) Detected MC positions obtained by thresholding *y*.

SVM Versus RVM

As explained earlier, SVM and RVM are both kernel methods, and both base the decision on only a subset of the training data—the support vectors in SVM and relevance vectors in RVM—that characterize the respective classes. However, SVM and RVM tend



[FIG4] (a) Comparison of support vectors from SVM and (b) relevance vectors from RVM for detection of MCs. SVM automatically chooses the support vectors to be examples lying near the decision boundary (hence the "MC absent" and "MC present" support vectors look very similar), while the relevance vectors chosen by RVM tend to be more prototypical of the two classes (hence the two groups of relevance vectors look very different).

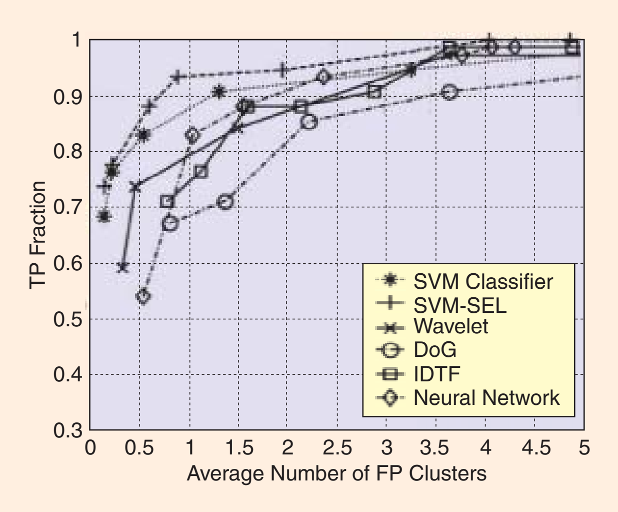
to select very different vectors to represent the classes. SVM chooses support vectors that lie very close to the decision boundary, while RVM tends to choose relevance vectors that are more prototypical of the two classes. Examples of support vectors and relevance vectors are shown in Figure 4. Note that the "MC present" and "MC absent" support vectors are very difficult to distinguish, as they all lie near the decision boundary, while the "MC present" and "MC absent" relevance vectors are clear examples of lesion and background regions, respectively.

ALTHOUGH RESAMPLING RECEIVES LESS ATTENTION THAN PREDICTIVE MODELS, IT IS AT LEAST AS IMPORTANT.

CADx: DIAGNOSIS OF CLUSTERED MICROCALCIFICATIONS

A great deal of research has been directed toward computerized CADx methods designed to assist radiologists in the difficult decision of differentiating benign from malignant MCs. In [15], a CADx scheme was demonstrated to classify clustered MCs even more accurately than radiologists. This method used a feedforward neural network (FFNN), which was trained using metrics extracted automatically from the clustered MC images.

Motivated by recent developments in machine learning, we sought in [16] to determine whether state-of-the-art machine-learning methods [SVM, kernel Fisher discriminant (KFD), RVM, and committee machines (including ensemble averaging and Adaboost, a well-known boosting method)] would further improve classification of MC clusters as malignant or benign, as compared with prior methods such as FFNN. We used the features defined in [15] that are based on both the shape and size of individual MCs as well as their overall distribution as a cluster, that are known to correlate qualitatively to features used by radiologists.



[FIG5] Detection performance of various methods of detecting MCs in mammograms. The best performance was obtained by a successive learning SVM classifier, which achieves around 94% detection rate (TP fraction) at a cost of one FP cluster per image, where a classical technique (DoG) achieves a detection rate of only about 68%.

The evaluation study demonstrated that the kernel methods (SVM, KFD, and RVM) are similar in performance to one another (in terms of the area under the receiver-operating characteristic (ROC) curve), but all demonstrated statistically significant improvement over FFNN or AdaBoost.

CBIR FOR CADx

Though promising, CADx has met with resistance to adoption in clinical practice, in part because radiologists are trained to interpret visual data and rarely deal with quantitative mammographic information, such as the likelihood of malignancy. Thus, when presented with a numerical value, but without additional supporting evidence, it may be difficult for a radiologist optimally to incorporate this number into the diagnostic decision. As such, traditional CADx classifiers are often criticized for being a "black box" approach.

To avoid this pitfall, an alternative approach we have advocated is to employ CBIR [17], [18], in which an image search engine is used to inform the radiologist's diagnosis in difficult cases by presenting relevant information from past cases. The retrieved example lesions allow the radiologist to explicitly compare known cases to the unknown case. A key advantage of this approach is that it provides case-based evidence to support case-based reasoning by the radiologist, rather than acting as a supplemental decision maker.

For a retrieval system to be useful as a diagnostic aid, the retrieved images must be truly relevant to the query image as perceived by the radiologist, who otherwise may simply dismiss them. In 2000 [17], we proposed a supervised learning approach for modeling the radiologists' notion of image similarity for use in CBIR. Our rationale is that mathematical distance metrics designed for general-purpose image retrieval may not adequately characterize clinical notions of image relevance, which are complex assessments made by expert observers.

In our approach, the perceptual similarity between two lesion images is modeled by a nonlinear regression model applied to the image features. The model is determined by using supervised learning from examples collected either in human observer studies or from online user feedback (acquired during use of the system). Specifically, we first characterize a lesion by vector $\mathbf{u}$ containing its key relevant features. Next, feature vector $\mathbf{u}$ is compared to the corresponding feature vector $\mathbf{v}$ of a database entry by way of predictive model $f(\mathbf{u}, \mathbf{v})$ to produce a similarity coefficient (SC). The images with the highest SC values are retrieved from the database and displayed for the user. In our studies, we have modeled $f(\mathbf{u}, \mathbf{v})$ using a nonlinear regression SVM and a general regression neural network (GRNN). Our learning metric has proven to be much more effective than alternative measures [17], [18].

To illustrate perceptual similarity, Figure 6 is a plot created using a multidimensional scaling (MDS) algorithm showing 30 microcalcification clusters. MDS is a family of techniques that aim to map high-dimensional data into a lower-dimensional representation in such as a way as to preserve relative distances (i.e., if two points are close to one another in the high-dimensional space, then MDS attempts to place them near one another in the low-dimensional space).

FOR A RETRIEVAL SYSTEM TO BE USEFUL AS A DIAGNOSTIC AID, THE RETRIEVED IMAGES MUST BE TRULY RELEVANT TO THE QUERY IMAGE AS PERCEIVED BY THE RADIOLOGIST, WHO OTHERWISE MAY SIMPLY DISMISS THEM.

In Figure 6, each microcalcification cluster is represented by a marker (square or circle) in the scatter plot. MDS attempts to place the points so that visually similar microcalcification clusters (as judged by human observers) are placed close to one another in the scatter plot. Examples of the microcalcfication clusters corresponding to these data points are shown as collections of plus (+) signs. Visual inspection of these examples suggests that the vertical axis of the plot is associated roughly with density of the microcalcifications, while the horizontal axis reflects the shape of the cluster. Note that there is a reasonable, but not perfect, separation between malignant and benign lesion classes in this space.

Recently, we proposed to use CBIR to boost the performance of a traditional CADx classifier [18]. Specifically, database images similar to the image being evaluated by the radiologist are used to improve the SVM classifier, thus improving its accuracy in analyzing the present case. We are currently investigating the impact of CBIR on the diagnostic performance of radiologists.

AUTOMATED ASSESSMENT OF IMAGE QUALITY BY PREDICTION OF DIAGNOSTIC PERFORMANCE

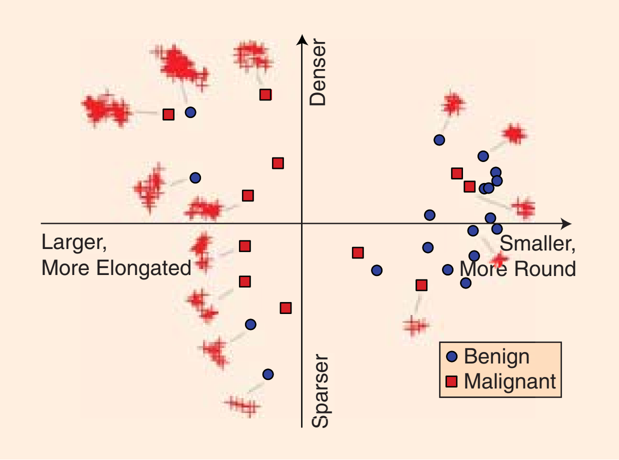
Diagnostic imaging can be thought of as a pipeline consisting of an imaging device, an image processor (e.g., image reconstruction algorithm and display), and a human observer (e.g., a radiologist). Principled methods are needed to assess the impact of design choices in the image acquisition and processing stages on the final interpretation stage.

It has been common traditionally to evaluate imaging devices and image reconstruction software using only basic fidelity metrics, such as signal-to-noise ratio (SNR), mean-square error, and bias and variance. However, such metrics have limitations when comparing images affected by statistically different types of blur, noise, and artifacts [19]. This was recognized in the 1970s in the context of radiographic imaging by Lusted [20], who pointed out that the image can reproduce the shape and texture of tissues faithfully from a physical standpoint, while failing to contain useful diagnostic information. In a highly influential article in *Science* [20], Lusted postulated that, to measure the worth of a diagnostic imaging test, one must assess the observer's performance when using the imaging test. In other words, if an image is to be used for lesion detection, then image quality should ideally be judged by the ability of an observer to detect lesions. Such an approach has become known as task-based assessment of image quality.

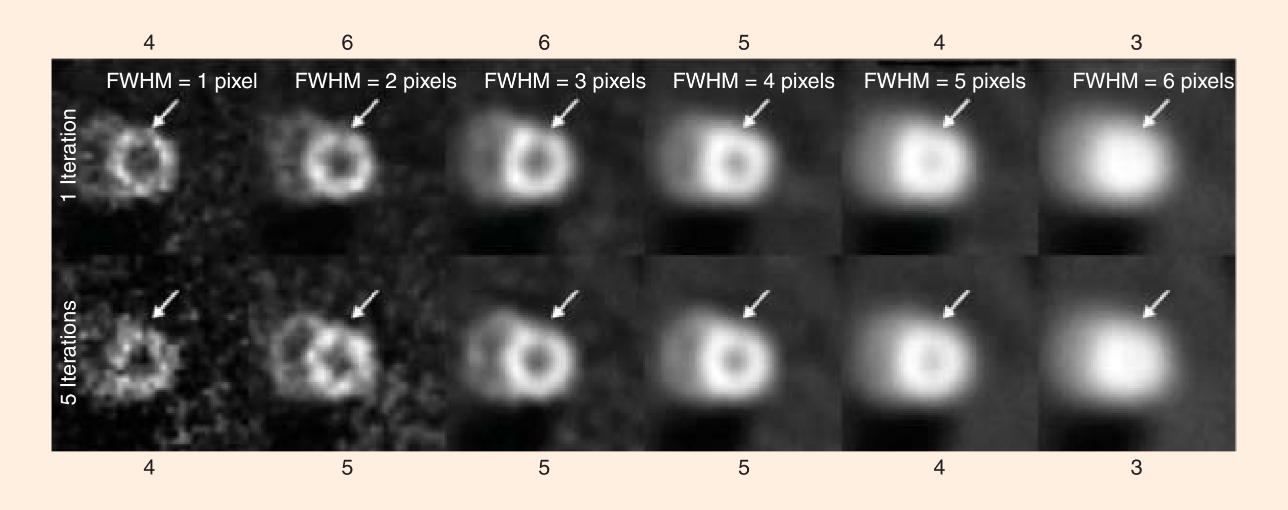
Lusted further argued that the ROC curve from classical detection theory is an ideal means to characterize diagnostic performance, and thus image quality. This approach has led to the wide use of ROC analysis in medical imaging, as implemented, for example, in the ROCKIT software distributed by Metz et al. [21].

Figure 7 shows an example of how the human observer's performance is affected by the type of images that are presented. In this case, the observer is shown a perfusion image of the myocardium (heart wall), obtained using single-photon emission computed tomography (SPECT). The observer is asked to judge whether there is a dark region indicating deficient perfusion, based on images reconstructed in different ways from the very same data set. Figure 7 shows 12 different reconstructions obtained by using either one or five iterations of the ordered-subset expectation-maximization algorithm (OS-EM), and with Gaussian filters having varying full width at half-maximum (FWHM).

Along the top and bottom of Figure 7 are values of an observer's stated confidence in the presence of a lesion at a location indicated by arrows (on a scale of one to six, with six indicating high confidence). Note that the observer's confidence



[FIG6] Statistical tool for visualizing relationships among abnormalities seen in various mammograms, in which distances reflect the relative similarities of abnormalities, as judged by human experts. MC clusters are represented in this two-dimensional diagram by using multidimensional scaling, a statistical technique that seeks to represent high-dimensional data in a lower-dimensional plot that can be readily visualized, while aiming to maintain the relative distances (similarities) among the data points. Each group of red plus signs (+) depicts the actual MC cluster associated with a given point in the scatter plot. This shows that the vertical axis of the plot is roughly associated with the density of each cluster, while the horizontal axis is related to its shape.



[FIG7] A human observer's judgment as to the presence of an abnormality (in this case a cardiac perfusion defect) depends on the parameters of the reconstruction algorithm used to create the image (here, the parameters are number of iterations and width (FWHM) of the post-reconstruction smoothing kernel). All of the images above have a defect at the location indicated by the arrow, but persons asked to judge whether there is a defect varied in their opinions from a value of three, meaning "defect is possibly not present," to a value of six, meaning "defect is definitely present." Our algorithm's ability to predict this behavior permits us to optimize a given algorithm for this specific diagnostic task.

that a lesion is present increases, then decreases, as the images are made smoother. Selection of the optimal smoothing level is an example of a goal in which a quantitative image-quality metric is needed.

MACHINE-LEARNING MODEL OF HUMAN OBSERVERS

In diagnostic imaging, the gold standard for measuring image quality is a statistical study that measures observers' (e.g., radiologists') diagnostic performance when using a given set of images. Unfortunately, the expense and complexity of such studies precludes their routine use. Therefore, numerical observers—algorithms that emulate human observer performance—are now widely used as surrogates for human observers.

One particular numerical observer, known as the channelized Hotelling observer (CHO) [22], has come to be widely used, particularly in nuclear medicine imaging. The CHO is a Fisher LD applied to input features obtained by applying band-pass (channel) filters to the image. These channels are inspired by the notion of receptive fields in the human visual system. Because of its principled approach to image quality evaluation, the CHO has justifiably had a major and positive impact on the field and has enjoyed tremendous popularity.

However, the CHO does not perfectly capture human-observer performance; therefore, we have proposed a new approach in which the problem of task-based image-quality assessment is viewed as a supervised-learning or system-identification problem [23]. That is, the goal is to identify the unknown human observer mapping, $f(\mathbf{x})$, between the image features in $\mathbf{x}$ and an observer score y that reflects the human observer's confidence in the presence of an abnormality in the image. This relationship is learned from example data obtained from human observers; the model is then used to make predictions in new situations where no human-observer data are available.

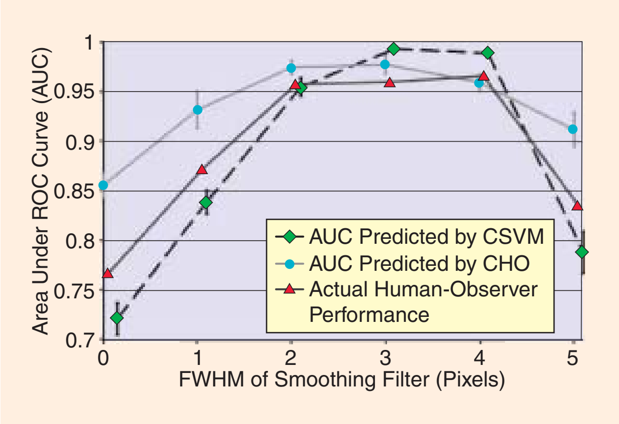
In our work, we have thus far retained the channels used in the CHO, contained in vector $\mathbf{x}$, but we feed these as inputs to a SVM $f(\mathbf{x})$, which we train to predict observer score y based on training examples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$. The resulting algorithm is called a channelized SVM (CSVM).

RESULTS

In [23], we compared the CSVM to the CHO for assessment of image quality in cardiac SPECT imaging. In this experiment, two medical physicists evaluated the defect visibility in 100 noisy images and scored their confidence of a lesion being present on a six-point scale, following a training session involving an additional 60 images. The human observers performed this task for six different choices of the smoothing filter and two different choices of the number of iterations in the OS-EM reconstruction algorithm (see Figure 7).

To demonstrate the generalization power of this approach, we trained both the CHO and CSVM on a broad range of images, then tested both on a different, but equally broad, range of images. Specifically, we trained both numerical observers using images for every value of the filter FWHM and five iterations of OS-EM and then tested the observers using all the images for every value of the filter FWHM with one iteration of OS-EM. The parameters of the CHO and CSVM were fully optimized to minimize generalization error measured using five-fold cross validation based on the training images only. Therefore, no test images were used in any way in the choices of the model parameters for either numerical observer. The numerical observers' predictions of human observers' area under the ROC curve (AUC) are compared in Figure 8 to human observers' actual performance. In this situation, the CHO performed relatively poorly, failing to match either the shape or amplitude of the human-observer AUC curves, while the CSVM was able to produce reasonably accurate predictions of AUC in both cases. Each error bar represents the standard deviation calculated using five-fold cross validation on the testing data.

This experiment demonstrates the potential benefit of using machine learning to make predictions rather than fixed models. Owing to the generality of its approach, machine learning can be used to make predictions of human-observer performance in many clinical tasks other than lesion detection, while CHO is specifically designed for lesion detection and is therefore less amenable to generalization.



[FIG8] **Predictions of human-observer performance (AUC) by machine learning approach (CSVM) compared with conventional numerical observer (CHO). The CHO does not recognize the degree to which diagnostic performance declines at low and high levels of smoothing, an effect seen in scores along the top and bottom of Figure 7.**

MAPPING OF BRAIN FUNCTION

Brain mapping is concerned with the creation of spatial representations (maps) of the brain, shedding light on the roles of various brain regions in normal and disease processes. Brain mapping is an area of application that differs significantly from those we have discussed thus far in the following two principal respects: 1) in many situations, brain mapping is concerned less with the prediction outputs y than with the model $f(\mathbf{x})$ itself, from which brain maps are obtained; and 2) owing to the relatively small number of data examples available in brain mapping, nonlinear models are not always preferred over simpler linear methods.

Brain mapping has been a rapidly growing field of imaging for at least 25 years. It is impossible to give a balanced survey of this field and its use of machine learning in the space available, so we will give only a brief overview.

In the 1980s, brain mapping was dominated by positron emission tomography (PET) and SPECT. The first machine-learning approaches to the analysis of functional brain images applied artificial neural networks (ANNs) to PET images of glucose metabolism [24]. However, following the discovery of the blood oxygenation level dependent (BOLD) signal in 1990 that allows regional neuronal activity to be measured indirectly, there has been explosive growth in the use of functional magnetic resonance imaging (fMRI) and related techniques [25].

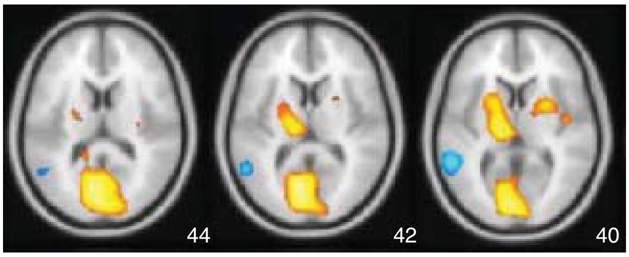
The prevailing experimental and analysis paradigm in brain mapping is still based on simple, univariate general linear models (GLM) with inferential statistical tests [26], and in some instances their predictive, machine-learning equivalent, Gaussian Naïve Bayes [27]. There has been a recent surge of papers and interest in using related multivariate classification approaches, dubbed "mind reading" by some in the field. For recent reviews including a historical perspective see [28], and for an overview of the often overlooked power of simple multivariate approaches, e.g., principal component analysis and LD, applied to PET scans of disease groups, see [29], which reflects the results of more than 20 years of work on measuring covariance structures that reflect brain networks. This network theme has gained considerable momentum in the more recent fMRI brain mapping literature with a focus on measuring the so-called "default mode" brain network using pair-wise, voxel correlations [30], or seed-voxel/behavioral partial least squares (PLS) [31], independent component analysis (ICA) [32], [33], and most recently nonlinear dynamics [34] and graph theory coupled with structural scans of white-matter networks [35].

Much of our own work has focused on the question of how to evaluate and optimize performance, and how to select the best signal detector from the broad repertoire of machine learning tools available. We have particularly focused on the impact of smaller sample sizes where analytic asymptotic theory for multivariate machine learning models, if it exists, does not provide much, if any, guidance. Analysis of brain images is a highly ill-posed problem, in which there are typically tens or hundreds of thousands of voxels, but only tens or hundreds of brain scans. Therefore, this small sample limit is the most likely to be important for medical use in brain mapping.

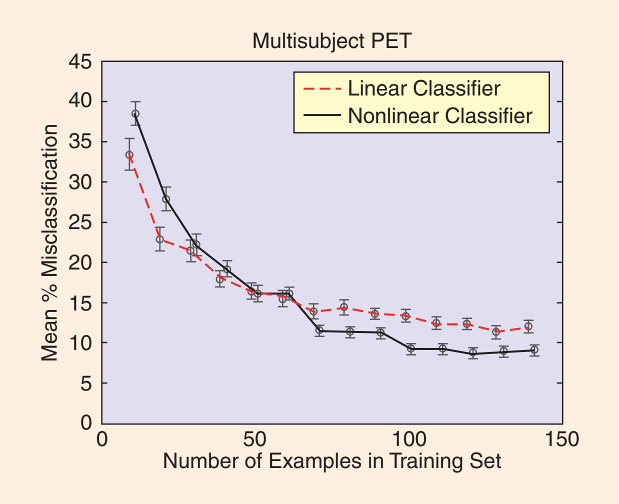
DISCRIMINANT IMAGES AS BRAIN MAPS

To illustrate the use of machine learning in brain mapping, let us consider one type of study in which we wish to produce an image showing the regional effects of a new drug on brain function (two of the authors of this article, Wernick and Strother, conduct such analyses commercially for the pharmaceutical industry). To accomplish this, one can scan a group of N research subjects twice, once after the subject is given the new drug and once after administration of placebo. One can then analyze these $2N$ images to obtain an image that describes the drug's effect. It is hoped that this finding will describe not only this particular group of subjects but will also generalize to some broader population.

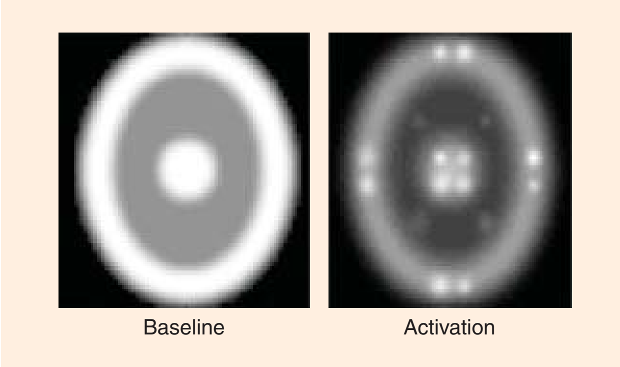
The basic idea underlying many machine-learning approaches to this problem is to treat each image as a vector in a high-dimensional space, with each component representing the value of one voxel in a scan. In this example, our data can be viewed as consisting of two classes of images: drug and placebo. To reduce dimensionality to a manageable level, and to mitigate noise, it is



[FIG9] Spatial activation pattern in the brain, showing effect of the anxiolytic/antidespressant drug buspirone (Buspar) obtained using Fisher LD and NPAIRS split-half resampling applied to FDG-PET images for 12 subjects (data courtesy of Abiant, Inc.; analysis by Predictek, Inc.). The results show striatal activation (upper orange regions), likely due to the drug's behavior as a dopamine D2 receptor antagonist.



[FIG10] These crossed learning curves (plots of classifier performance versus training set size) show that a nonlinear classifier (a neural network in this example) can be beaten by a simpler multivariate linear classifier (here, a Fisher discriminant) when the number of training examples is small. This is not unexpected, as small data sets cannot generally support complex models, however this result emphasizes the importance of resisting the temptation for researchers to use high-complexity models in every circumstance.



[FIG11] Simulated phantom used for testing signal detection.

common to transform the data using singular value decomposition (SVD). Next, a classifier is trained to discriminate drug images from placebo images based on the dimensionality-reduced data.

In traditional pattern classification applications, the purpose of training the classifier is to make decisions about new data. Indeed, there are a growing number of examples of this in neuroimaging, for example in lie detection, or in diagnosis of disease in an individual patient. However, in many studies, the goal is simply to understand what intrinsically is different about the brain in, say, a drug and a placebo condition. In such instances, the desired information is encoded in the predictive model $f(\mathrm{x})$ itself. When a linear model is used, then the desired brain map is encoded in the components of discriminant vector w, which (after projecting back from SVD space to image space) describes the salience of voxels in the brain for discrimination of drug and placebo conditions.

Figure 9 shows an example of such an image (which we will refer to as a spatial activation pattern) after it is thresholded and overlaid on a template structural image used to bring multiple subjects' brains into an approximate common space. The value of each colored voxel in this image expresses the degree to which that voxel contributes to the discrimination of drug versus placebo, and this image thereby depicts the spatial distribution of effect.

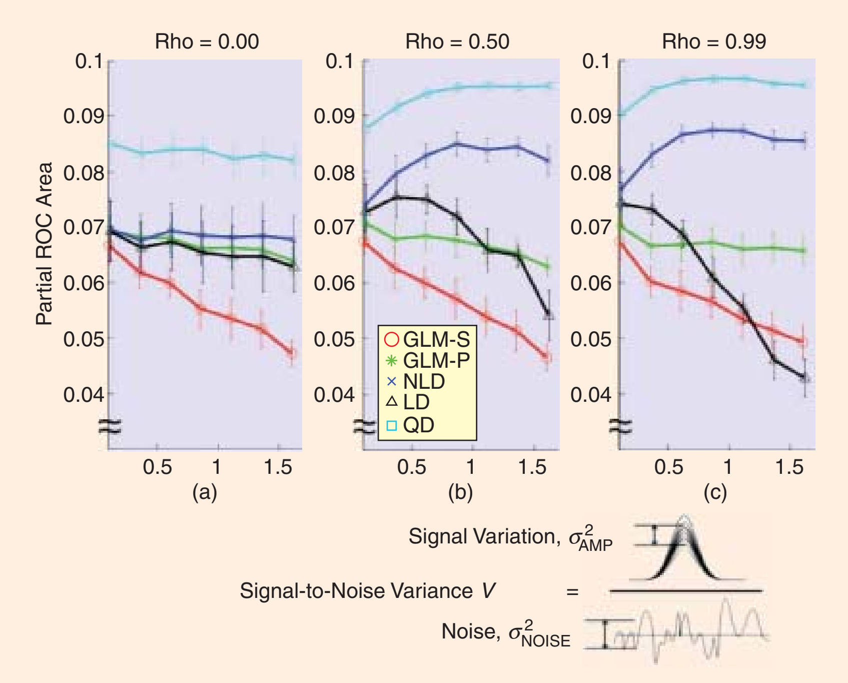
Note that, in this basic introduction, we have refrained from describing a significant series of preprocessing steps that must be applied before the machine learning algorithms can be used. These are discussed at length in [36].

COMPARING MODELS, SAMPLE SIZE, AND SNR

Evaluations of data-analysis techniques have clearly illustrated that optimal tool selection depends critically on the signal and noise structure of the data at hand, and the sample size [37], [38]. For example, Figure 10 (adapted from [38]) illustrates that a simple linear model can outperform a flexible nonlinear model (in this case an ANN) until there are enough data examples to support estimation of the greater number of parameters inherent in the nonlinear model. Nevertheless, these issues are frequently ignored in the current brain mapping literature when discussing or comparing different analysis techniques.

We have addressed the question of choosing optimal analysis procedures using simulations in [39] based on the simple phantom shown in Figure 11, assuming an experimental design similar to the drug-placebo study described earlier. We varied numerous parameters of the simulation, including number of examples per condition (from 20 to 100), and the amplitude of the activation "blobs" in the phantom (either 3% or 5% above baseline). We added spatially colored, temporally white, Gaussian noise with a standard deviation of 5% of the mean baseline value. We created three spatially distributed "networks" of blobs, and varied the correlation coefficient ρ (rho) between them ($\rho = 0.0, 0.5$, or 0.9) and the ratio V of their amplitude variance to the noise variance. This ratio can be thought of in analogy to dynamic range in audio, as the blob variance is a source of signal in this application, which is of particular relevance for the field's recent focus on network detection in brain mapping. In [39], we showed that SVD by itself or followed by a LD that adapts the subspace on which it is estimated is much more sensitive to network interactions than thresholding of pairwise correlation coefficients [40].

We have repeated and extended the earlier work of Lukic et al. using the same phantom (results shown in Figure 12). Simulations included 3% Gaussian amplitudes, with 30 baseline and 30 activation scans. The models tested include 1) single-voxel t-tests using both local (GLM-S) and spatially pooled (GLM-P) variance estimates, and classification techniques including a 2) two-class Fisher LD, 3) normalized LD (NLD), and 4) quadratic discriminant (QD). All multivariate techniques were estimated on an SVD subspace with dimension determined using optimization of Bayes' evidence [41], as estimated in the software package MELODIC [42]. For LD and QD, the SVD basis components had length equal to their eigenvalues, and for NLD they were normalized to unit length.



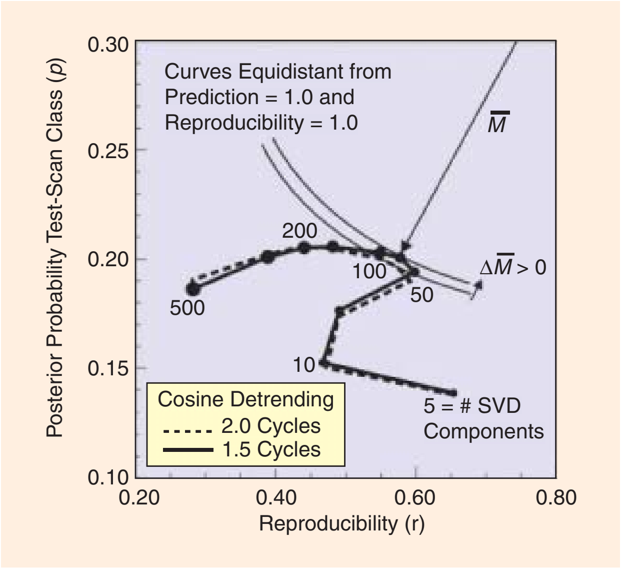
[FIG12] In (a)–(c), performance in detection of brain activation for five models, as a function of signal-to-noise variance ratio (V) and correlations (rho) among network of activated brain regions, are shown. The QD and NLD perform best, improving with strength of network (increasing V and rho), while the performance of univariate methods lags behind, and actually deteriorates as the signal strength increases.

Using the area under the ROC curve for false positives between [0.0, 0.1], signal detection was measured across the 16 voxels at the peaks of the Gaussian blobs. Even when the t-test with local variance estimates (GLM-S) was the "correct" model (i.e., $V = 0.1$) better detection performance was obtained using a t-test with a pooled variance estimate or adaptive, multivariate covariance-based detectors. In addition, GLM-S showed a significant drop in performance as the equal variance assumption was violated with increasing V. Variance estimation by spatially pooling (GLM-P) significantly improved signal detection and largely removed this source of model violation.

The multivariate equivalent of the GLM-S model violation is shown by the LD results where the assumption of equal within-class covariances (i.e., a common network structure for baseline and activation scans) is violated with increasing V; only the activation scans have an off-diagonal, within-class covariance structure that increases with V. However, LD still outperforms GLM-S for all but the strongest violations of the equal covariance assumption for large rho and V [Figure 12(c)]. In the NLD method, the standard machine-learning trick of normalizing input feature variances (i.e., unit SVD basis vectors) significantly improves signal detection performance to always better than GLM-P, and largely removes the LD drop with increasing V. Finally, using the correct multivariate model that assumes different within-class covariances, a QD, further significantly improves performance to close to perfect (partial ROC area approaches 0.1). QD, as used here, represents an alternative to SVM as a solution to the problem of unequal class distributions shown in Figure 2.

The relative performance of LDs and SVM remains controversial in brain mapping with some papers claiming SVM is superior [43] and others that they are approximately equal [44], but that they respond to different input SNR structures differently as suggested by the analysis of Figure 2. Moreover, our most recent simulation results show that signal detection performance is a very strong function of the SVD basis set size and performance may be improved even further than shown in Figure 11 by using a resampled estimate of the optimal SVD subspace based on the reproducibility metric outlined below.

Our final simulation results relate to a comparison of Bayesian kernel methods with a generalized likelihood ratio test for estimating local activation in functional neuroimages. In [45], we compared spatial signal detection using the superposition of spatial Gaussian kernels with their parameters estimated from the data using a maximum a posteriori (MAP) technique based on a reversible-jump Markov-chain Monte Carlo (RJMCMC) algorithm and a RVM. RVM and RJMCMC were better signal detectors than all of the other techniques tried in [39] and achieved values of 0.80 and 0.82 for the partial area under the ROC curve. These performance values cannot be directly compared to Figure 11 as the simulation parameters were quite different. However, the RJMCMC took tens of hours to compute, even in our simple phantom, while the RVM was computed in



[FIG13] In the NPAIRS framework, a prediction-reproducibility (p,r) curve shows the tradeoff between prediction accuracy (vertical axis) and reproducibility of the resulting brain map (horizontal axis). Optimal performance is achieved when the curve comes closest to the ideal point (1,1), achieving the smallest distance $\overline{M}$. This provides a basis for optimizing image analysis procedures, in this example specifying the best parameters in a particular fMRI data analysis problem (number of SVD components and number of cycles in a particular cosine detrending step).

only minutes. The relative utility of SVM, RVM, and other kernel techniques in brain mapping (e.g., kernel PCA, [28]; kernel canonical correlation analysis [46]) remains to be established.

DATA-DRIVEN PERFORMANCE METRICS

In brain mapping, as in general machine-learning applications, it is very important to optimize and evaluate predictive models and to select their most salient features. These tasks must be guided by a quantitative metric of performance. Prediction accuracy often plays this role, for example to guide a greedy search procedure to select the most salient subset of voxels [26]. Some tradeoffs of such purely prediction-driven analysis approaches are discussed in [4] and [27].

Although prediction accuracy alone can be an effective metric for general machine-learning problems, neuroimaging also demands that the spatial pattern (encoded by the predictive model) be reproducibile between different groups of subjects or different scans of the same subject. Together with prediction accuracy, reproducibility turns out to be an important metric that is a very effective data-driven substitute for ROC analysis.

Strother et al. [9] proposed a novel split-half resampling framework dubbed NPAIRS, which simultaneously assesses prediction accuracy and reproducibility. The tradeoff between achievable prediction accuracy and reproducibility of the model is related to the classic tradeoff of bias and variance in estimation theory. In this application, prediction accuracy is generally gained at the expense of decreased reproducibility of the spatial patterns, and vice versa. By plotting prediction accuracy versus reproducibility as a function of some parameter (such as number of SVD basis vectors), we are able to assess the gamut of this tradeoff, in close analogy to the ROC curve, the precision-recall curve from the information retrieval field, or the bias-variance curve from statistics. We call this type of plot produced by the NPAIRS analysis a (p, r) curve.

To compute a (p, r) curve using NPAIRS, the independent observations of the data set are split into two independent halves (e.g., across subjects): training and test sets. Prediction accuracy is obtained by applying the spatial patterns estimated in one split-half set (i.e., training) to estimate scan class labels in the other split-half set (i.e., test). The roles of the two split-half sets are then reversed so that the each set has been used once as a training set (to produce a spatial activation pattern) and once as a test set. From these results, two prediction accuracy estimates (p) are computed and averaged to obtain the overall prediction accuracy. Next, the reproducibility of the two independent spatial activation patterns is computed as the correlation (r) between all pairs of spatially aligned voxels in the two patterns. This correlation value r is directly related to the available SNR in each extracted pair of split-half patterns. If one forms a scatter plot consisting of the voxel values in one spatial pattern versus corresponding values in the other, one obtains a distribution in which the principal, or signal, axis has associated eigenvalue $(1+r)$, and the uncorrelated minor, or noise, axis has eigenvalue $(1-r)$. Therefore, one can define a global data set SNR metric gSNR as

$$\text{gSNR} = \sqrt{((1+r)-(1-r))/(1-r)} = \sqrt{2r/(1-r)}.$$

In NPAIRS, many split-half resamplings are performed and the average, or median, of the resulting p and r distributions are recorded. This resampling approach has the benefits of smooth robust metrics obtained with the 0.632+ bootstrap [8]. Finally, a robust consensus technique is used to combine the many split-half spatial patterns into a single pattern described on a Z-score (standard normal) scale, providing a robust Z-scoring mechanism for any prediction model that produces voxel-based parameter estimates.

In [29], NPAIRS was applied to PET, and it has been also been applied to fMRI [47]–[49]. While NPAIRS may be applied to any analysis model, we have particularly focused on LDs, and more recently QDs, both built on an SVD basis. This allows us to 1) regularize the model by choosing soft (e.g., ridge) or hard thresholds on an SVD or other basis set [50], 2) maintain the link to covariance decomposition that has proven so useful in PET for elucidating network structures, and 3) produce whole-brain activation maps that enhance the likelihood of discovering new features of brain function and disease.

Figure 13 shows an example of how NPAIRS can be used to study the influence of the key parameters of an image analysis procedure, and thus permit one to make an optimal selection of these parameters. In this example, two parameters of an fMRI image analysis procedure are examined, the number of SVD basis vectors (defining model complexity) and the number of half cosines used for detrending [36]. (We will not elaborate here on details of the SVD and detrending techniques; we show this example only to illustrate how NPAIRS can in general be used to select optimal model parameters.)

In a (p, r) plot, ideal performance is achieved by reaching the upper right corner of the space, where prediction accuracy (described as posterior probability in Figure 13) reaches 1.0 and reproducibility also achieves 1.0. Thus, one approach to defining the optimal choice of parameters is to determine the point at which the (p, r) curve attains the least Euclidean distance $(\overline{M})$ to the point (1,1). In this example, we see that performance [distance to (1,1)] improves, then worsens, as the number of SVD components increases. The effect of the cosine detrending parameter is weaker, but indicates that one and a half cycles is a somewhat better choice than two cycles. In this graph, the hook-shaped portion between five and ten SVD components represents reproducible artifacts that are commonplace in fMRI.

The NPAIRS analysis framework provides a very useful way to understand and optimize model performance in the challenging problem of brain mapping, and perhaps in other applications in which one is interested not only in making accurate predictions but also in producing reliable information on the factors driving these predictions.

ACKNOWLEDGMENTS

The authors wish to acknowledge numerous collaborators who contributed to the research summarized in this article, including Nikolas P. Galatsanos, Lars Kai Hansen, Issam El-Naqa, Ana S. Lukic, Robert M. Nishikawa, Stephen LaConte, David Rottenberg, Liyang Wei, and Jane Zhang.

The research reviewed in this article was sponsored in part by NIH/NCI grant CA89668, NIH/NIBIB grant R01EB009905, NIH/NIBIB grant HL091017, NIH/NINDS grant NS34069, NIH/NIMH grant MH073204, James S. McDonnell Foundation Ph.D. scholarship to Grigori Yourganov, NIH/NIBIB P20EB02013, NIH/NIMH P20MH072580, and CIHR/MOP84483. Stephen C. Strother gratefully acknowledges support of the Heart & Stroke Foundation of Ontario through the Centre for Stroke Recovery.

AUTHORS

Miles N. Wernick (wernick@iit.edu) received the B.A. degree in physics from Northwestern University in 1983 and the Ph.D. degree in optics from the University of Rochester in 1990. In 1990, he was an NIH Postdoctoral Fellow in radiology at the University of Chicago, where he became a research associate assistant professor. In 1994, he joined the Illinois Institute of Technology, where he is currently director of the Medical Imaging Research Center and Motorola Endowed Chair Professor of Engineering in the Departments of Electrical and Computer Engineering and Biomedical Engineering. He is also president of Predictek, Inc. His research interests include medical imaging, machine learning, image processing, and optics. He is guest editor of this special issue of *IEEE Signal Processing Magazine*, an associate editor of *IEEE Transactions on Image Processing* and *SPIE/IS&T Journal of Electronic Imaging*, and a member of the IEEE Bioimaging and Signal Processing Technical Council.

Yongyi Yang (yy@ece.iit.edu) received the B.S.E.E. and M.S.E.E. degrees from Northern Jiaotong University, Beijing, China, in 1985 and 1988, respectively. He received the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the Illinois Institute of Technology (IIT), Chicago, in 1992 and 1994, respectively. He is currently a professor in the Department of Electrical and Computer Engineering at IIT, where he is with the Medical Imaging Research Center and also holds a joint appointment with the Department of Biomedical Engineering. His research interests are in signal and image processing, medical imaging, machine learning, pattern recognition, and biomedical applications. He is an associate editor of *IEEE Transactions on Image Processing*.

Jovan G. Brankov (brankov@iit.edu) received the diploma of electrical engineering from the University of Belgrade, Yugoslavia, in 1996. He received the M.S.E.E. and Ph.D. degrees in electrical engineering from the IIT in 1999 and 2002, respectively. He is currently an assistant professor in the Department of Electrical and Computer Engineering at IIT, where he is with the Medical Imaging Research Center. His research interests include medical imaging, image sequence processing, pattern recognition, and data mining. His current research topics include four-dimensional and five-dimensional tomographic image reconstruction methods for medical image sequences, multiple-image radiography (a new phase-sensitive imaging method), and image quality assessment based on a human-observer model. He is author/coauthor of over 80 publications and serves as ad hoc associate editor for *Medical Physics*.

Grigori Yourganov (gyourganov@rotman-baycrest.on.ca) received the B.S. and M.S. degrees in computer science from York University, Toronto, Canada, in 2000 and 2005, respectively. He is currently completing his Ph.D. degree in the Institute for Medical Sciences at Rotman Research Institute (University of Toronto), under the supervision of Dr. Stephen C. Strother and Dr. Randy McIntosh. His research is focused on application of multivariate analytical techniques to fMRI data.

Stephen C. Strother (sstrother@rotman-baycrest.on.ca) received the B.Sc. and M.Sc. degrees in physics and mathematics from Auckland University, New Zealand in 1976 and 1979, respectively, and a Ph.D. degree in electrical engineering from McGill University, Montreal in 1986. Since 1985, he has been a postdoctoral fellow at Memorial Sloan-Kettering Cancer Center, New York. In 1989 he joined the VA Medical Center, Minneapolis, as senior PET physicist, and the University of Minnesota where he became a professor of radiology in 2002. In 2004, he moved to Toronto as a senior scientist at the Rotman Research Institute and professor of medical biophysics at the University of Toronto, where he is also a core member of the multiinstitutional Centre for Stroke Recovery. His current research interests include neuroinformatics with a focus on machine and statistical learning techniques for optimizing PET and fMRI/MRI neuroimaging in research and clinical applications applied to the aging brain. In 2001 he cofounded Predictek, Inc., in Chicago. He is an associate editor for *Human Brain Mapping*.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2003.

[2] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2001, p. 626.

[3] M. N. Wernick, “Pattern classification by convex analysis,” *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 8, pp. 1874–1880, 1991.

[4] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[5] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sept. 2001.

[6] R. G. Baraniuk, E. J. Candès, R. Nowak, and M. Vitterli, “Compressive sampling,” *IEEE Signal Processing Mag.*, vol. 21, no. 2, pp. 12–13, Mar. 2008.

[7] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: CRC, 1994.

[8] B. Efron and R. Tibshirani, “Improvements on cross-validation: The .632+ bootstrap method,” *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 548–560, June 1997.

[9] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg, “The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework,” *Neuroimage*, vol. 15, no. 4, pp. 747–771, Apr. 2002.

[10] *Image-Processing Techniques for Tumor Detection*. New York: Marcel Dekker, 2002.

[11] *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*. Bellingham, WA: SPIE, 2006.

[12] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, “Computer-aided detection and diagnosis of breast cancer with mammography: recent advances,” *IEEE Trans. Inform. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Mar. 2009.

[13] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, “A support vector machine approach for detection of microcalcifications,” *IEEE Trans. Med. Imaging*, vol. 21, no. 12, pp. 1552–1563, Dec. 2002.

[14] L. Wei, Y. Yang, R. M. Nishikawa, M. N. Wernick, and A. Edwards, “Relevance vector machine for automatic detection of clustered microcalcifications,” *IEEE Trans. Med. Imaging*, vol. 24, no. 10, pp. 1278–1285, Oct. 2005.

[15] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, “Malignant and benign clustered microcalcifications: automated feature analysis and classification,” *Radiology*, vol. 198, no. 3, pp. 671–678, Mar. 1996.

[16] L. Wei, Y. Yang, R. M. Nishikawa and Y. Jiang, “A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications,” *IEEE Trans. Med. Imaging*, vol. 24, no. 3, pp. 371–380, Mar. 2005.

[17] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, “A similarity learning approach to content-based image retrieval: application to digital mammography,” *IEEE Trans. Med. Imaging*, vol. 23, no. 10, pp. 1233–1244, Oct. 2004.

[18] L. Y. Wei, Y. Y. Yang, M. N. Wernick, and R. M. Nishikawa, “Learning of perceptual similarity from expert readers for mammogram retrieval,” *IEEE J. Select. Topics Signal Processing*, vol. 3, no. 1, pp. 53–61, Feb. 2009.

[19] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, “Image quality assessment based on a degradation model,” *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[20] L. B. Lusted, “Signal detectability and medical decision making,” *Science*, vol. 171, pp. 1217–1219, 1971.

[21] C. E. Metz, B. A. Herman, and J. H. Shen, “Maximum-likelihood estimation of ROC curves from continuously-distributed data,” *Stat. Med.*, vol. 17, no. 9, pp. 1033–1053, 1998.

[22] K. J. Myers and H. H. Barrett, “Addition of a channel mechanism to the ideal-observer model,” *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 12, pp. 2447–2457, Dec. 1987.

[23] J. G. Brankov, Y. Yang, L. Wei, I. El Naqa, and M. N. Wernick, “Learning a channelized observer for image quality assessment,” *IEEE Trans. Med. Imaging*, vol. 28, no. 7, pp. 991–999, July 2009.

[24] J. Kippenhan, W. Barker, S. Pascal, J. Nagel, amd R. Duara, “Evaluation of a neural network classifier for PET scans of normal and Alzheimers disease subjects,” *J. Nucl. Med.*, vol. 33, pp. 1459–1467, 1992.

[25] P. Bandettini, “Functional MRI today,” *Int. J. Psychophysiol.*, vol. 63, no. 2, pp. 138–145, Feb. 2007.

[26] K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. New York: Academic, 2006.

[27] F. Pereira, T. Mitchell, and M. Botvinick, “Machine learning classifiers and fMRI: A tutorial overview,” *Neuroimage*, vol. 45, no. 1 (Suppl.), pp. S199–S209, Mar. 2009.

[28] L. K. Hansen, “Multivariate strategies in functional magnetic resonance imaging,” *Brain Lang.*, vol. 102, no. 2, pp. 186–191, Aug. 2007.

[29] D. Eidelberg, “Metabolic brain networks in neurodegenerative disorders: A functional imaging approach,” *Trends Neurosci.*, vol. 32, no. 10, pp. 548–557, Oct. 2009.

[30] M. D. Fox and M. E. Raichle, “Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging,” *Nat. Rev. Neurosci.*, vol. 8, no. 9, pp. 700–711, Sept. 2007.

[31] A. R. McIntosh, W. K. Chau, and A. B. Protzner, “Spatiotemporal analysis of event-related fMRI data using partial least squares,” *Neuroimage*, vol. 23, no. 2, pp. 764–775, Oct. 2004.

[32] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith, “Investigations into resting-state connectivity using independent component analysis,” *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 360, no. 1457, pp. 1001–1013, May 2005.

[33] N. M. Correa, T. Adalı, Y.-O. Li, and V. D. Calhoun, “Canonical correlation analysis for data fusion and group inferences,” *IEEE Signal Processing Mag.*, vol. 27, no. 4, pp. 39–50, 2010.

[34] K. E. Stephan, L. M. Harrison, S. J. Kiebel, O. David, W. D. Penny, and K. J. Friston, “Dynamic causal models of neural system dynamics: Current state and future extensions,” *J. Biosci.*, vol. 32, no. 1, pp. 129–144, Jan. 2007.

[35] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann, “Predicting human resting-state functional connectivity from structural connectivity,” *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 6, pp. 2035–2040, Feb. 2009.

[36] S. C. Strother, “Evaluating fMRI preprocessing pipelines,” *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 2, pp. 27–41, Mar.–Apr. 2006.

[37] N. Lange, S. C. Strother, J. R. Anderson, F. A. Nielsen, A. P. Holmes, T. Kolenda, R. Savoy, and L. K. Hansen, “Plurality and resemblance in fMRI data analysis,” *Neuroimage*, vol. 10, no. 3, part 1, pp. 282–303, Sept. 1999.

[38] N. Morch, L. K. Hansen, S. C. Strother, C. Svarer, D. A. Rottenberg, B. Lautrup, R. Savoy, and O. B. Paulson, “Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover,” in *Information Processing in Medical Imaging* (Lecture Notes in Computer Science), J. Duncan and I. Gindi, Eds. 1997, pp. 259–270.

[39] A. S. Lukic, M. N. Wernick, and S. C. Strother, “An evaluation of methods for detecting brain activations from functional neuroimages,” *Artif. Intell. Med.*, vol. 25, no. 1, pp. 69–88, May 2002.

[40] K. J. Worsley, J. Cao, T. Paus, M. Petrides, and A. C. Evans, “Applications of random field theory to functional connectivity,” *Hum. Brain Mapp.*, vol. 6, no. 5–6, pp. 364–367, 1998.

[41] T. P. Minka, “Automatic choice of dimensionality for PCA,” Cambridge, MA: MIT, Rep. 514, 2004.

[42] C. F. Beckmann and S. M. Smith, “Probabilistic independent component analysis for functional magnetic resonance imaging,” *IEEE Trans. Med Imaging*, vol. 23, no. 2, pp. 137–152, Feb. 2004.

[43] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, “Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data,” *Neuroimage*, vol. 28, no. 4, pp. 980–995, Dec. 2005.

[44] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, “Support vector machines for temporal classification of block design fMRI data,” *Neuroimage*, vol. 26, no. 2, pp. 317–329, June 2005.

[45] A. S. Lukic, M. N. Wernick, D. G. Tzikas, X. Chen, A. Likas, N. P. Galatsanos, Y. Yang, F. Zhao, and S. C. Strother, “Bayesian kernel methods for analysis of functional neuroimages,” *IEEE Trans. Med Imaging*, vol. 26, no. 12, pp. 1613–1624, Dec. 2007.

[46] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, “Unsupervised analysis of fMRI data using kernel canonical correlation,” *Neuroimage*, vol. 37, no. 4, pp. 1250–1259, Oct. 2007.

[47] S. C. Strother, S. La Conte, L. Kai Hansen, J. Anderson, J. Zhang, S. Pulapura, and D. Rottenberg, “Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis,” *Neuroimage*, vol. 23 (Suppl. 1), pp. S196–S207, 2004.

[48] J. Zhang, J. R. Anderson, L. Liang, S. K. Pulapura, L. Gatewood, D. A. Rottenberg, and S. C. Strother, “Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA,” *Magn. Reson. Imaging*, vol. 27, no. 2, pp. 264–278, Feb. 2009.

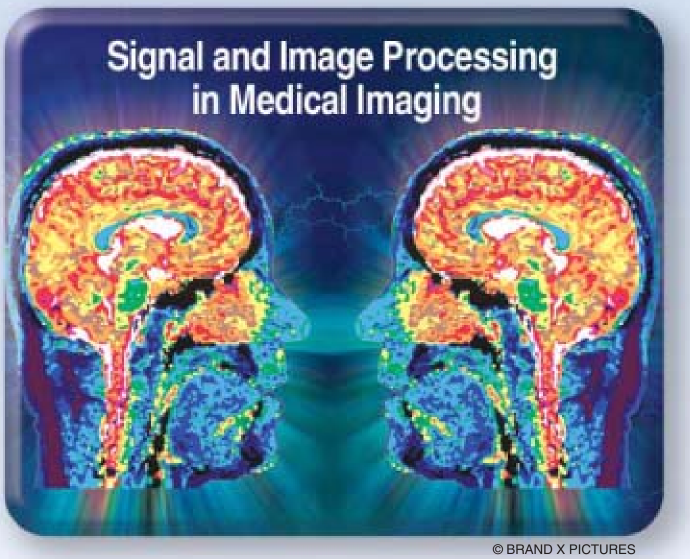
[49] J. Zhang, L. Liang, J. R. Anderson, L. Gatewood, D. A. Rottenberg, and S. C. Strother, “Evaluation and comparison of GLM- and CVA-based fMRI processing pipelines with Java-based fMRI processing pipeline evaluation system,” *Neuroimage*, vol. 41, pp. 1242–1252, July 2008.

[50] R. Kustra and S. C. Strother, “Penalized discriminant analysis of [15O]-water PET brain images with prediction error selection of smoothness and regularization hyperparameters,” *IEEE Trans. Med. Imaging*, vol. 20, no. 5, pp. 376–387, May 2001.

[Nicolle M. Correa, Tülay Adalı, Yi-Ou Li, and Vince D. Calhoun]

Canonical Correlation Analysis for Data Fusion and Group Inferences

[Examining applications of medical imaging data]



© BRAND X PICTURES

Data-driven analysis methods, such as blind source separation (BSS) based on independent component analysis (ICA), have proven very useful in the study of brain function, in particular when the dynamics are hard to model and underlying assumptions about the data have to be minimized. Many problems in medical data analysis involve the analysis of multiple data sets, either of the same type as in a group study where inferences are based on the same modality, e.g., group inferences from functional magnetic resonance imaging (fMRI) data collected from multiple subjects, or from different modalities as in the case of data fusion where inferences have to be drawn from data collected from multiple modalities such as fMRI, electroencephalography (EEG), and structural MRI (sMRI), for the same group of subjects. Canonical correlation analysis (CCA) [1], another data-driven approach, and its extension to multiple data sets—multiset CCA (M-CCA) [2]—provide a natural framework for both types of study. In this article, we show how CCA and M-CCA can be used for the analysis of data from a single modality for group inferences as well as fusion of data from multiple modalities using a feature-based approach, discuss the advantages of the CCA-based approach, and compare its performance to ICA that has been successfully applied to both types of study.

BACKGROUND

Analysis of multiple sets of data, either of the same type as in multitask or multisubject data, or of different type or nature as in multimodality data, is inherent to many fields and is a particularly challenging problem in biomedical image analysis because of the rich nature of the data made available by different imaging modalities. Analysis of multiple sets of same type of data is an integral part of biomedical imaging studies, e.g., when estimating brain activations in fMRI data from a group of subjects or when analyzing data from two different experimental conditions such as fMRI data from subjects scanned at different alcohol levels while performing a given task. Fusion of data from different modalities promises to provide a better understanding of the problem at hand since each modality has its own advantages as well as limitations. In the case of biomedical imaging, an increasing number of studies are collecting multiple measurements, e.g., fMRI data, sMRI data, EEG data,

Digital Object Identifier 10.1109/MSP.2010.936725

1053-5888/10/$26.00©2010IEEE

genetic data, and others from the same participants. Efficient use of all this information for inference, while minimizing assumptions made about the underlying nature of the data and relationships, is an arduous task but is one that promises significant gains in understanding of the human brain function. The main purpose of analyzing multiple modalities is to utilize the common as well as unique information from complementary modalities to better understand neuronal activity. For example, the fusion of fMRI data and EEG data—fMRI having good spatial resolution and EEG having high temporal resolution but poor spatial localization of brain activity—provides a better spatio-temporal mapping of the brain function. In this article, we address both types of problems: fusion of data sets collected from multiple modalities and the analysis of multisubject data from the same modality.

[MANY PROBLEMS IN MEDICAL DATA ANALYSIS INVOLVE THE ANALYSIS OF MULTIPLE DATA SETS, EITHER OF THE SAME TYPE AS IN A GROUP STUDY WHERE INFERENCES ARE BASED ON THE SAME MODALITY.]

Approaches to solve these multidata set problems can be broadly classified as being either model based or data driven. Model-based approaches investigate the goodness-of-fit of the data to the prior knowledge about the experimental paradigm and the properties of the data, for example the general linear model approach [3] for the analysis of fMRI data utilizes the prior knowledge of the hemodynamic properties of the data and the task. While model-based approaches have been extensively used in biomedical data analysis, their use is limited when the dynamics of the experiment become hard to model, e.g., when studying rest state or naturalistic paradigms such as driving or watching a movie. Data-driven methods are suitable for the analysis of such complex paradigms as they minimize the assumptions on the underlying properties of the data by decomposing the observed data based on a generative model. The most common decomposition is given by $\mathrm{X} = \mathrm{AS}$ (with the possibility of including an additive noise term), where X is the mixture that is factorized into latent variables through two matrices—a mixing matrix A and a component (source) matrix S. For uniqueness of the decomposition (subject to scaling and permutation ambiguity), constraints are applied to the two matrices such as sparsity or independence of the components. Model-based approaches provide a similar decomposition, however they differ from data-driven methods in their modeling of the matrix A, which is based on prior knowledge of the experiment and data in the form of regressors. ICA is a popular data-driven BSS technique that imposes the constraint of statistical independence on the components, i.e., source distributions. It has been successfully applied to a number of biomedical data such as fMRI [4], [5] and EEG [6]. A number of approaches have been proposed to solve the ICA problem, a popular one being the maximum likelihood estimation technique that finds an approximation of the underlying sources by using the maximum likelihood estimator of the demixing matrix W such that $\hat{\mathrm{S}} = \mathrm{WX}$. Second-order data-driven methods have also been used for biomedical data analysis such as linear discriminant analysis, partial least squares, CCA, and source-separation algorithms such as the Molgedey Schuster algorithm [7]. CCA has been used to find latent sources in single subject fMRI data by taking advantage of the spatial or temporal autocorrelation in the data [8]. An extension of the Molgedey Schuster algorithm has also been used to extract sources from two or more data sets based on the temporal autocorrelation in fMRI data [9]. In this article, we focus on reviewing CCA and M-CCA methods for data fusion and multisubject analysis of biomedical data and putting these into perspective through comparisons with closely related ICA-based methods.

DATA FUSION

Data-driven fusion of multimodality data is an especially challenging problem since brain imaging data types are intrinsically dissimilar in nature, making it difficult to analyze them together without making a number of assumptions, most often unrealistic about the nature of the data. Unlike data integration methods, which tend to use information from one modality to improve the other, data fusion techniques incorporate both modalities in a combined analysis, thus allowing for true interaction between the different data types [10]. Instead of entering the entire data sets into a combined analysis, an alternate approach is to reduce each modality to a feature, which is a lower-dimensional representation of selected brain activity or structure, and then to explore associations across these feature data sets through variations across individuals. Investigating variations across subjects or between patients and controls at the feature-level provides a natural way to find multimodality associations [11] and also alleviates the difficulty of fusing data types of different dimensionality and nature as well as those that have not been recorded simultaneously. Feature-level analysis has been successfully used in data-driven fusion techniques such as joint-ICA (jICA) [11] and CCA-based fusion [12]. Given two feature data sets X_1 and X_2, the jICA approach involves concatenating the data sets alongside each other and then performing ICA on the concatenated data set as in $[\mathrm{X}_1\mathrm{X}_2] = \mathrm{A}[\mathrm{S}_1\mathrm{S}_2]$. Joint-ICA assumes that the sources have a common modulation profile A across subjects, which is a strong constraint considering that the data come from two different modalities. Parallel-ICA (paraICA) [13] is another ICA-based feature-level fusion approach that has been successful in identifying relationships between neuroimaging data types as well as between genetic and phenotypic data. The method performs separate ICA on the different modalities in parallel while enhancing intrinsic correlations across the modalities. While jICA requires the two modalities to have relatively similar dimensions, the two separate ICA in paraICA is more flexible in that it allows the modalities to have different dimensions. For a review of feature-based fusion methods using ICA and their application to biomedical imaging, refer to [14] and [15].

Recently, it has been shown that CCA and M-CCA can allow a more flexible approach to the fusion problem. The CCA-based fusion method also adopts a feature-based approach and similarly models the feature data set from each modality as a linear mixture of components with varying levels of activations for different subjects. Thus, the relationship between modalities is based on intersubject covariations as shown in Figure 1(a). The scheme is flexible as the connections are based only on the linear mixing model and intersubject covariances across modalities, and the modulation profiles of components are not constrained to be exactly the same as in jICA. Successful application of CCA to feature-based fusion of two modalities has been shown in [12] and of M-CCA to the fusion of three brain imaging modalities has been shown in [16]. CCA and M-CCA have also been successfully used for fusion in other fields such as remote sensing [17] and pattern recognition [18].

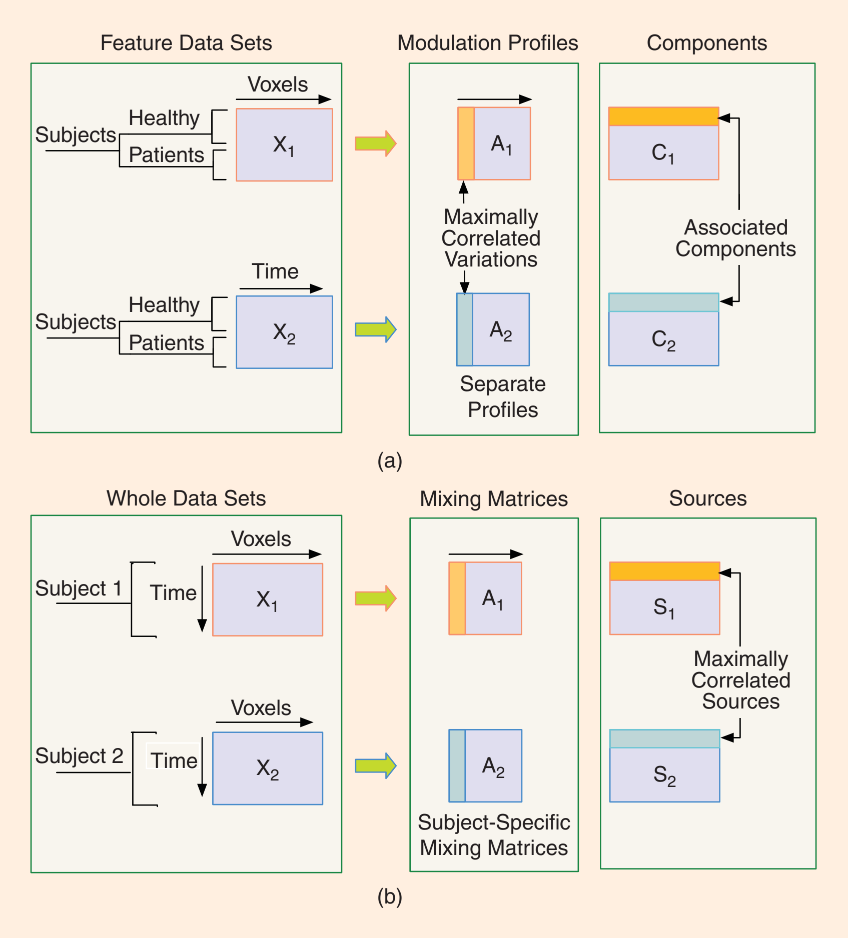
[THE MAIN PURPOSE OF ANALYZING MULTIPLE MODALITIES IS TO UTILIZE THE COMMON AS WELL AS UNIQUE INFORMATION FROM COMPLEMENTARY MODALITIES TO BETTER UNDERSTAND NEURONAL ACTIVITY.]

MULTISUBJECT ANALYSIS

M-CCA can be used to perform group BSS, i.e., source separation of single modality data from multiple subjects [19]. While it is straightforward to apply data-driven techniques such as BSS to each subject's data separately, the challenge lies in matching the separated sources across different data sets, which is straightforward in model-based methods. M-CCA provides an effective tool to perform group BSS while maintaining the correspondence of the source estimates across different data sets and retaining the intersubject source variability. The generative model for M-CCA for BSS is shown in Figure 1(b). A number of data-driven methods have been proposed for achieving group BSS and can be broadly categorized into two different approaches. One approach is to concatenate multiple data sets to aggregate the common features, perform analysis in the common feature space to estimate group components, and back-project the estimated group components into each data set to obtain individual components with cross-data set correspondence. Group ICA [20] and tensorial ICA [21] fall into this category. The other approach assumes a generative model on latent components with cross-data set correspondence and performs group component extraction using statistical measures of correspondence. M-CCA and independent vector analysis (IVA) [22] fall into this category; however, M-CCA and IVA are complementary in modeling component correspondence [19]. Compared with methods based on data set concatenation, M-CCA is more flexible in identifying cross-data set variation of the components, which can be used for making group level inferences in different ways. M-CCA has been shown to be successful for the group analysis of fMRI data in [19].

OUTLINE

In this article, we begin with a description of the main statistical tools for fusion and multisubject analysis—CCA and M-CCA. We then provide a brief introduction to the medical imaging data that we will be using to demonstrate CCA- and ICA-based analysis techniques. There are two formats in which the data will be utilized by the approaches we present: 1) the whole multidimensional data, e.g., for fMRI data, both the time and the volume information contained in all image slices of the brain, and 2)



[FIG1] Generative models for fusion and source separation are shown in (a) and (b). To avoid overfitting, typically dimension reduced data matrices are used instead of the high-dimensioned data X_1 and X_2. For data fusion, the spatial or temporal dimensions are reduced. For group analysis, the temporal dimension is reduced.

lower-dimensional features extracted from the whole data to represent certain aspects of the data, e.g., for fMRI data, areas of the brain where neuronal activity due to task is exhibited excluding the time information for these areas. The presented fusion approaches are carried out at the feature-level using CCA, M-CCA, or jICA while the group analysis is carried out on whole multidimensional data sets using M-CCA and group-ICA. To further illustrate the use of the methods, we provide a number of examples of medical imaging applications for the presented techniques.

CCA

CCA has been traditionally used to analyze relationships between two sets of variables [1]. CCA seeks two sets of transformed variates such that the transformed variates assume maximum correlation across the two data sets, while the transformation within each data set are uncorrelated. CCA is an attractive analysis tool, based on second-order statistics and is less stringent than those based on stronger statistical measures such as ICA. Also, being multivariate, it can provide increased statistical power over univariate methods.

Given two data sets $\mathbf{X}_1 \in \mathbb{R}^{n\times p}$ and $\mathbf{X}_2 \in \mathbb{R}^{n\times q}$, CCA finds the linear combinations $\mathbf{X}_1\mathbf{W}_1$ and $\mathbf{X}_2\mathbf{W}_2$ that maximize the pair-wise correlations across the two data sets. $\mathbf{A}_1$ and $\mathbf{A}_2 \in \mathbb{R}^{n\times d}$, $d \le \min(\mathrm{rank}(\mathbf{X}_1, \mathbf{X}_2))$, are known as canonical variates and $\mathbf{W}_1 \in \mathbb{R}^{p\times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{q\times d}$ are the canonical coefficients vectors.

In the deflationary approach, the method finds the first pair of canonical coefficient vectors $\mathbf{w}_1^{(1)}$ and $\mathbf{w}_2^{(1)}$, $(\mathbf{w}_1^{(1)} \in \mathbb{R}^{p\times 1}$, $\mathbf{w}_2^{(1)} \in \mathbb{R}^{q\times 1})$ that maximize linear combinations of the two data sets given by

$$\max_{\mathbf{w}_1^{(1)},\,\mathbf{w}_2^{(1)}} \mathrm{corr}(\mathbf{X}_1\mathbf{w}_1^{(1)}, \mathbf{X}_2\mathbf{w}_2^{(1)})$$

to obtain the first pair of canonical variates given by

$$\mathbf{a}_1^{(1)} = \mathbf{X}_1\mathbf{w}_1^{(1)} \text{ and } \mathbf{a}_2^{(1)} = \mathbf{X}_2\mathbf{w}_2^{(1)}.$$

The remaining $d-1$ canonical variates can be calculated similarly, with the following additional constraints on the columns of the A matrices, i.e., $\mathbf{a}_k^{(i)}$ $(i = 1, \ldots, d,\ k = 1, 2)$:

- The canonical variates are uncorrelated within each data set and have zero mean and unit variance, i.e.,

$$\mathbf{A}_k^T\mathbf{A}_k = \mathbf{I}, \quad k = 1, 2. \tag{1}$$

- The canonical variates have nonzero correlation only on their corresponding indices, and have correlation coefficients, $r_{k,l}^{(1)} \ge r_{k,l}^{(2)}, \ldots, \ge r_{k,l}^{(d)}$, where $r_{k,l}^{(i)} = \mathbf{a}_k^{(i)T}\mathbf{a}_l^{(i)}$, i.e.,

$$\mathbf{A}_k^T\mathbf{A}_l = \mathbf{R}_{k,l}, \quad k \ne l, k, l = 1, 2, \tag{2}$$

where $\mathbf{R}_{k,l} = \mathrm{diag}(r_{k,l}^{(1)}, \ldots, r_{k,l}^{(d)})$.

The CCA problem can be posed as a constrained optimization problem using Lagrange multipliers and the canonical covariates can be calculated by solving a generalized eigenvalue solution, where the columns of $\mathbf{W}_1$ and $\mathbf{W}_2$ are the eigenvectors of the two matrices $\mathbf{C}_{X_1}^{-1}\mathbf{C}_{X_1,X_2}\mathbf{C}_{X_2}^{-1}\mathbf{C}_{X_2,X_1}$ and $\mathbf{C}_{X_2}^{-1}\mathbf{C}_{X_2,X_1}\mathbf{C}_{X_1}^{-1}\mathbf{C}_{X_1,X_2}$, where $\mathbf{C}_{X_1,X_2}$ is the cross-correlation matrix of $\mathbf{X}_1$ and $\mathbf{X}_2$ $(\mathbf{C}_{X_2,X_1} = \mathbf{C}_{X_1,X_2}^T)$, and $\mathbf{C}_{X_1}$ and $\mathbf{C}_{X_2}$ are the autocorrelation matrices of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

M-CCA

The CCA problem can be extended to multiple data sets using the framework developed in [2]. In contrast to CCA where correlation between two canonical variates is maximized, M-CCA optimizes an objective function of the correlation matrix of the canonical variates from multiple random vectors such that the canonical variates achieve maximum overall correlation. Furthermore, due to the consideration of multiple random vectors, M-CCA can not be solved by a simple eigenvalue decomposition problem as in the case of CCA. Instead, M-CCA takes multiple stages such that in each stage, one group of canonical variates is obtained by optimizing the objective function with respect to a set of transformation vectors. For the second stage and higher stages in M-CCA, the estimated canonical variates are constrained to be uncorrelated to the ones estimated in the previous stages. M-CCA reduces to CCA when the number of random vectors is two. Given K data sets, the canonical variates

$$\mathbf{A}_k = \mathbf{X}_k\mathbf{W}_k, \quad k = 1, 2, \ldots, K \tag{3}$$

can be estimated through a deflationary approach such that we first determine the initial K vectors corresponding to the first source from each of the K data sets using

$$\{\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_K^{(1)}\} = \arg\max_{\mathbf{w}} J(r_{k,l}^{(1)})$$

and then the next vectors using the same procedure such that $\mathbf{w}_k^{(i)}$ is orthogonal to the previous estimates, i.e., to $\{\mathbf{w}_k^{(1)}, \mathbf{w}_k^{(2)}, \ldots, \mathbf{w}_k^{(i-1)}\}$, $k = 1, 2, \ldots, K$. Here, $r_{k,l}^{(i)}$ is the correlation between the ith canonical variates, from the kth and lth data sets, estimated in the final decomposition and $J(\cdot)$ is an appropriately chosen cost. The canonical correlations can be obtained by optimizing a number of cost functions proposed in [2], e.g., maximizing the sum of squared correlations among the canonical variates.

We can summarize the M-CCA procedure based on the sum of squares correlation (SSQCOR) cost as

- Stage 1

$$\{\mathbf{w}_1^{(1)}, \mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_K^{(1)}\} = \arg\max_{\mathbf{w}}\left\{\sum_{k,l=1}^{K} |r_{k,l}^{(1)}|^2\right\} \tag{4}$$

- Stage 2 to d

```
for $i = 2{:}d$
    $\{\mathbf{w}_1^{(i)}, \mathbf{w}_2^{(i)}, \ldots, \mathbf{w}_K^{(i)}\} = \arg\max_{\mathbf{w}}\left\{\sum_{k,l=1}^{K} |r_{k,l}^{(i)}|^2\right\}$
    s.t. $\mathbf{w}_k^{(i)} \perp \{\mathbf{w}_k^{(1)}, \mathbf{w}_k^{(2)}, \ldots, \mathbf{w}_k^{(i-1)}\}, k = 1, 2, \ldots, K$
end
```

For M-CCA , up to d canonical variates can be calculated iteratively, where $d \le$ $\min(\mathrm{rank}(\mathbf{X}_k))$. In [2], Stage 1 is solved by first calculating the partial derivative function of the SSQCOR cost with respect to each $\mathbf{w}_k^{(1)}$ and equating it to zero to find the stationary point. Since the SSQCOR cost is a quadratic function of each $\mathbf{w}_k^{(1)}$, the partial derivative is a linear function of $\mathbf{w}_k^{(1)}$ and hence, the closed-form solution can be derived. Starting from an initial point, each $\mathbf{w}_k^{(1)}$ vector is updated in sequel to guarantee an increase in the cost function and a sweep through all the $\mathbf{w}_k^{(1)}$ constitutes one step of the iterative maximization procedure. The iterations are stopped when the cost convergence criterion is met and the resulting $\mathbf{w}_k^{(1)}$ vectors are taken as the optimal solution. Stage 2 and higher stages are solved in a similar manner with the cost function replaced by a Lagrangian incorporating the orthogonality constraints on the canonical coefficient vectors.

Next, we explain the biomedical imaging modalities, the generative models, and some examples of the application of CCA and M-CCA for data fusion and source-separation applications.

M-CCA PROVIDES AN EFFECTIVE TOOL TO PERFORM GROUP BSS WHILE MAINTAINING THE CORRESPONDENCE OF THE SOURCE ESTIMATES ACROSS DIFFERENT DATA SETS AND RETAINING THE INTERSUBJECT SOURCE VARIABILITY.

MEDICAL IMAGING MODALITIES AND FEATURE GENERATION

In this article, we demonstrate the effectiveness of the CCA-based approach using three modalities: fMRI, sMRI, and EEG. Each of these modalities provides limited information about the human brain. FMRI is a noninvasive brain imaging technique that provides information about brain function by measuring the changes in blood-oxygenation in the brain. SMRI provides information about the tissue type of the brain—gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). EEG records brain function by measuring the brain electrical field through the scalp. For fusion, we adopt a feature-level analysis to derive a lower dimensional feature from the imaging data as in [14] and [20]. A feature is a subdata set extracted from one type of data, related to a selected brain activity or structure. These features can then be analyzed to integrate or fuse the information across multiple modalities. Next, we briefly introduce the data types, the preprocessing used for each of these data types, and the types of features we generate for the fusion analysis from each of the data sets.

fMRI

FMRI data provide a measure of brain function on a millimeter spatial scale and a subsecond (and delayed) temporal scale. The data consists of repeatedly imaging the 3-D volume of the brain slice-by-slice, usually while the subject performs a particular task. A number of preprocessing are steps important for fMRI—slice-timing correction to correct for the sequential acquisition of the slices, registration to correct for subject motion in the scanner, spatial filtering to reduce noise, and spatial normalization to compare brains across different individuals and to use standardized atlases to identify particular brain regions. For fMRI data, we use the task-related spatial activity map as calculated by the GLM approach as the spatial feature for the fusion analysis.

sMRI

We define sMRI analysis as the acquisition and processing of T1-, T2-, and/or proton density-weighted images. Multiple structural images are often collected to enable multispectral segmentation approaches. The primary outcome measure in a structural image may include a measure of a particular structure (e.g., volume or surface area) or a description of the tissue type, (e.g., GM or WM). There are many methods for preprocessing sMRI data that may include bias field correction [intensity changes caused by radio frequency (RF) or main magnetic field (B_0) inhomogeneities] [23], spatial linear or nonlinear [24] filtering, and normalization. MR images are typically segmented using a tissue classifier producing images showing the spatial distribution of GM, WM, and CSF. Both supervised and automated segmentation approaches have been developed for sMRI analysis [25]–[27], and each technique is optimized to detect specific features. We use probabilistically segmented GM images as features of sMRI data for the fusion analysis.

EEG

EEG is a technique that measures brain function by recording and analyzing the scalp electrical activity generated by brain structures. Like MRI, it is a noninvasive procedure that can be applied repeatedly in patients, normal adults and children, with virtually no risks or limitations. Local current flows are produced when brain cells are activated. It is believed that contributions are primarily driven by large synchronous populations of firing neurons. The recorded electrical signals are then amplified, digitized, and stored.

Event-related potentials (ERPs) are small voltage fluctuations resulting from evoked neural activity and are one of many ways to process EEG data. These electrical changes are extracted from scalp recordings by computer averaging epochs (recording periods) of EEG time locked to repeated occurrences of sensory, cognitive, or motor events. The spontaneous background EEG fluctuations, which are typically random relative to when the stimuli occurred, are averaged out, leaving the event-related brain potentials. These electrical signals reflect only that activity that is consistently associated with the stimulus processing in a time-locked way. The ERP thus reflects, with high temporal resolution, the patterns of neuronal activity evoked by a stimulus. Due to their high temporal resolution, ERPs provide unique and important timing information about brain processing and are an ideal methodology for studying the timing aspects of both normal and abnormal cognitive processes. More recently, ICA has been used to take advantage of EEG activity that may be averaged out by computing an ERP [6]. For the feature-based fusion analysis we use ERPs as the EEG feature.

DATA FUSION

In this section, we present the data fusion scheme at the feature level using CCA and M-CCA. We explain the data generation model, the modeling assumptions for feature-based fusion using CCA methods, and the dimension reduction step that is used to avoid overfitting. Additionally, we demonstrate the use of the method by presenting two examples and compare the CCA-based methods with ICA-based fusion method, jICA.

THE ANALYSIS OF MORE THAN TWO BRAIN IMAGING MODALITIES COLLECTIVELY, E.G., fMRI, sMRI, AND EEG, CAN HELP IDENTIFY INTERESTING ASSOCIATIONS ACROSS BRAIN STRUCTURE AND FUNCTION.

GENERATIVE MODEL FOR DATA FUSION

We develop the following generative model for data fusion. Given two feature data sets X_1 and X_2, we seek to decompose them into two sets of components, C_1 and C_2, and corresponding modulation profiles (intersubject variations), A_1 and A_2 as shown in Figure 1(a). The connection across the two modalities can be evaluated based on correlations of modulation profiles of one modality with those of the other. If the modulation profiles are uncorrelated within each modality, each component can be associated with only one component across modalities. This one-to-one correspondence aids in the examination of associations across modalities. The generative model is thus given by

$$X_k = A_k C_k, \quad \text{for} \quad k = 1, 2,$$

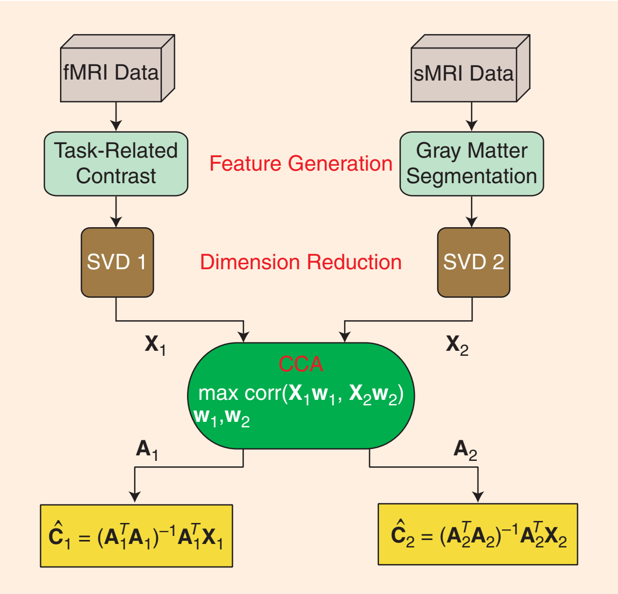
where $X_k \in \mathbb{R}^{n \times v_k}$, $A_k \in \mathbb{R}^{n \times d}$, $C_k \in \mathbb{R}^{d \times v_k}$, v_k is the number of variables in X_k, n is the number of observations in X_k, and $d \leq \min[\text{rank}(X_1, X_2)]$. The modeling assumptions imply that the modulation profiles, given by columns of A_1 and A_2 satisfy the constraints given by (1) and (2). In the feature-based fusion approach [12], the intersubject covariations across the two modalities, i.e., the correlations across the modulation profiles are identified using CCA as described in the section "CCA." The feature-based fusion scheme models the modulation profiles A_1 and A_2 as the canonical variates obtained by CCA, and based on the modulation profiles identified, the associated components can be calculated using least squares approximations given by

$$\hat{C}_k = (A_k^T A_k)^{-1} A_k^T X_k, \quad \text{for} \quad k = 1, 2.$$

Thus, this fusion approach identifies the cross-modality covariations, and based on these, it decomposes each feature data set into a set of components—such as spatial areas for fMRI/sMRI or temporal segments for EEG.

Typically, the number of variables (voxels/time points) in the feature data sets is much larger than the number of observations (subjects). Due to the high dimensionality and high noise levels in the brain imaging data, order selection is critical to avoid overfitting the data. Transforming each set of features to a subspace with smaller number of variables helps reduce any redundancy in the analysis. The dimension is chosen to fall in a range where the results are stable and most of the variance in the data can be retained. Dimension reduction is performed on the feature data set using singular value decomposition (SVD), and we perform CCA on the dimension-reduced data sets. We assume a noiseless generative model since we perform dimension reduction. i.e., the assumption in the SVD-based dimension reduction scheme is that small singular values of the matrix that are discarded correspond to additive noise.

The generative model we have described with respect to two data sets can be extended to multiple data sets. For example, for three data sets the CCA fusion method again models the modulation profiles A_1, A_2, and A_3 as the canonical variates—however, it is worth noting that in this case the canonical variates are obtained using M-CCA. The procedure for M-CCA is as described in the section "M-CCA" The calculation of the components as well as the dimension reduction steps for the features are the same as described above.



[FIG2] Implementation steps for CCA-based fusion.

COMPARISON WITH jICA

Joint-ICA has been successfully used for the fusion of data from two modalities such as fMRI, EEG, and sMRI data [11], [28]. The jICA approach is similar to the CCA-based fusion approach in that it is a second-level analysis based on lower-dimensional features of the data and the associations across the two modalities are based on intersubject covariations. However, there are a number of differences in the modeling assumptions of the two methods. Most importantly, jICA assumes that the sources share a common modulation profile while CCA-based fusion models the modulation profiles of each modality to be separate. Given the diverse nature of the two modalities, assuming that the modulations are exactly the same across different modalities can be a very strong constraint. Another important difference between the methods is that the associations across modalities in CCA-based fusion are solely based on intersubject covariations whereas the associations in jICA are based on the assumptions of common profiles as well as statistical independence among the joint sources. While CCA provides a relatively less constrained solution to the fusion problem, jICA utilizes higher-order statistical information by employing ICA. When correlations are strong between the two modalities, the assumption of a common mixing matrix may be justified and the jICA technique could potentially improve approximation of the joint sources by employing higher-order statistics in the estimation. However, by allowing for separate mixing matrices, CCA-based fusion promises to identify common as well as distinct components and reliably estimates the amount of association between the two modalities. For a detailed comparison of the two models as well as experimental results based on simulated fMRI-like and ERP-like data, refer to [12].

[FUSING INFORMATION FROM THE TWO MODALITIES COULD HELP TO UNDERSTAND THE LINK BETWEEN BRAIN STRUCTURE AND FUNCTION.]

APPLICATION OF CCA TO FUSION OF TWO MODALITIES

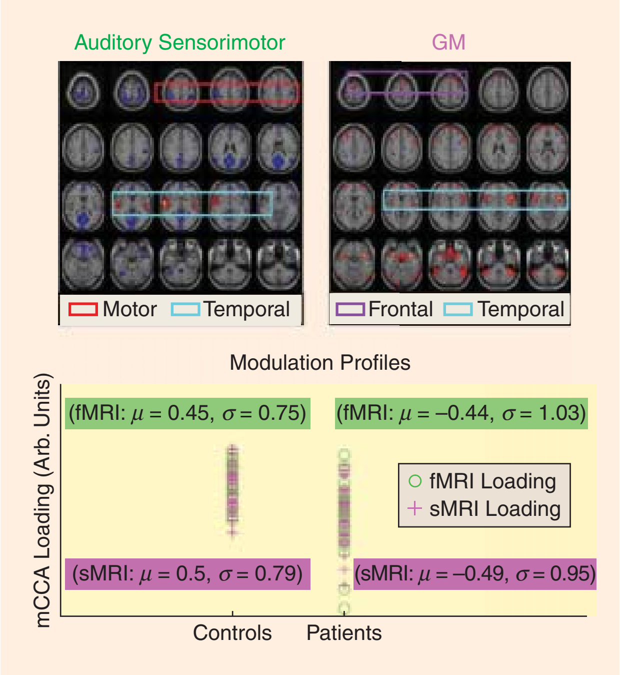
The CCA-based fusion approach has been successfully used to analyze the spatio-temporal associations between fMRI data and EEG data, and also, to detect functional and structural relationships between fMRI data and sMRI data in [12]. Here we discuss few of the key findings on the fusion of fMRI and sMRI data and compare the results to those obtained using the jICA method presented in [12].

FUSION OF fMRI AND sMRI

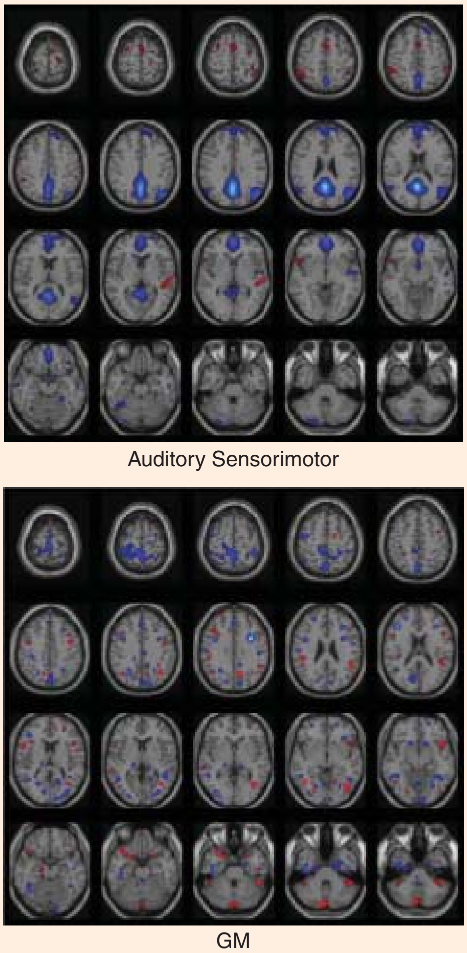
FMRI data provides information about brain function while sMRI data contains information about brain structure. Fusing information from the two modalities could help to understand the link between brain structure and function. We demonstrate the fusion approach on sMRI data and fMRI data from 37 patients with schizophrenia and 36 healthy controls carrying out an auditory sensorimotor task consisting of patterns of eight tones, alternately increasing and decreasing in pitch. The subjects are instructed to press a button with their right thumb for each presented tone. Details of the experimental setup are given in [29]. The fMRI data and sMRI data are converted into lower-dimensional features using the preprocessing techniques described in the section "Medical Imaging Modalities and Feature Generation." The reduced dimension for both features was empirically chosen as 18.

The pair of components corresponding to profiles showing the strongest correlation ($r^{(1)}_{1,2} = 0.87$) across the two data sets demonstrate significant group differences ($\alpha \leq 0.05$: $t_{\text{fMRI}} = -4.92$ and $t_{\text{sMRI}} = -4.80$) between patients with schizophrenia and healthy controls (fMRI map, sMRI map, and scatter plots of the profiles are shown in Figure 3). The fMRI component map shows that healthy controls have more functional activity in the temporal areas (activations enclosed in blue box) and less motor activity (activations enclosed in red box) compared to patients with schizophrenia. The GM map shows that healthy controls have more GM compared to the patients in frontal (activations enclosed in purple box) and temporal areas (activations enclosed in blue box). These results reveal associations between the modalities in adjacent or close sets of voxels as well as remotely located voxels. This is consistent with previous studies showing changes in both brain structure and brain function in frontal and temporal lobe regions in schizophrenia and is also in agreement with previous studies on fusion of fMRI and GM [28], [30], [31].

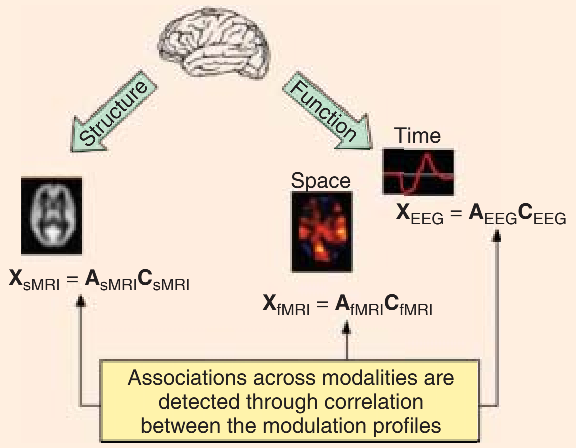
We also perform jICA on this data set using the fusion ICA toolbox (http://icatb.sourceforge.net/fusion/fusion_startup.php). The result obtained using jICA, shown in Figure 4, is similar to the one obtained using CCA (Figure 3). However, CCA shows additional motor and temporal areas. Also the structural regions are more localized and well defined in the CCA-based result. In general, however, the jICA components are mostly sparse, at least sparser than components obtained using CCA-based fusion, due to the non-Gaussian emphasis by ICA algorithms such as [32]. In


Auditory Sensorimotor
GM
Motor
Temporal
Frontal
Temporal
Modulation Profiles
mCCA Loading (Arb. Units)
(fMRI: μ = 0.45, σ = 0.75)
(fMRI: μ = –0.44, σ = 1.03)
○ fMRI Loading
+ sMRI Loading
(sMRI: μ = 0.5, σ = 0.79)
(sMRI: μ = –0.49, σ = 0.95)
Controls
Patients

[FIG3] The fMRI component, sMRI component, and scatter plots of profiles for pair of components identified by CCA as maximally correlated. The profiles for both fMRI and sMRI are significantly different ($\alpha \leq 0.05$) between patients and controls. Patients with schizophrenia show more functional activity in motor areas and less activity in temporal areas associated with less gray matter as compared to healthy controls. The activation maps are scaled to Z values and thresholded at $Z = 3.5$.



[FIG4] Joint components estimated by jICA corresponding to common profile demonstrating significant difference between patients and controls. The activation maps are scaled to Z values and thresholded at $Z = 3.5$.


Structure
Function
Time
Space
$X_{EEG} = A_{EEG}C_{EEG}$
$X_{sMRI} = A_{sMRI}C_{sMRI}$
$X_{fMRI} = A_{fMRI}C_{fMRI}$
Associations across modalities are detected through correlation between the modulation profiles

[FIG5] Data model for fusion of brain structure and function.

the fusion example presented here, this was seen for the fMRI result but not for the sMRI result. The sparseness of the sMRI results for CCA is hence interesting and may be due to the fact that CCA relaxes the strong constraint of common profiles for the pair of components.

APPLICATION OF M-CCA FOR FUSION OF MULTIPLE MODALITIES

A number of approaches have been proposed to integrate or fuse multitask or multimodality data. However, these have mostly been limited to two modalities or multiple data sets from the same modality. In [16], M-CCA was demonstrated to be successful in fusing data from three modalities. Next, we highlight the key findings from [16] and present new results that demonstrate the increase in sensitivity of the analysis with addition on more modalities.

FUSION OF fMRI, sMRI, AND EEG

The analysis of more than two brain imaging modalities collectively, e.g., fMRI, sMRI, and EEG, can help identify interesting associations across brain structure and function. Performing M-CCA on multiple data sets can be more restrictive since we are requiring covariation of all three modalities, however, this is also informative since we find changes that are related across the three modalities. An interesting point to note is that M-CCA-based fusion allows for associations in local voxels as well as remotely located voxels, thus enabling discoveries of structural changes causing compensatory functional activation in distant, but connected, regions.

Again, we decompose the data into sets of components and their corresponding modulation profiles across the subjects as shown in Figure 5. The data fusion scheme determines the linear transformation that maximizes the intersubject covariations across the three modalities using M-CCA, and based on these covariations, the associations among the components across modalities are determined. As an example for multimodality fusion using M-CCA, consider the fusion of three brain imaging modalities: fMRI, sMRI, and EEG. The MRI and EEG data are acquired from 36 subjects (22 healthy controls and 14 schizophrenia patients). The fMRI and EEG data were collected while the subjects performed an auditory oddball (AOD) task that required them to press a button when they detect a particular infrequent sound among three kinds of auditory stimuli. Details of the task design and the participants are given in [33]. The data was preprocessed and features were obtained as described in the section "Medical Imaging Modalities and Feature Generation." EEG features, or ERPs, are calculated from the midline central position (Cz) because it appeared to be the best single channel to detect both anterior and posterior sources for the given task.

We perform CCA on the dimension-reduced fMRI, sMRI, and ERP data to estimate 15 sets of components that contain interesting associations across the modalities. The results identify changes in the motor and temporal areas associated with the N2/P3 complex in the ERP (the EEG feature) as shown in Figure 6, areas that have been also previously noted as affected in schizophrenia. On examining the intersubject modulation in conjunction with the spatial and temporal components, the results imply that subjects with schizophrenia have less functional activity and less GM in the areas detected in this component and also a part of the ERP response appeared to be affected. Note that in the section "Application of CCA to Fusion of Two Modalities," the sensorimotor task showed an increase in motor activity for patients with schizophrenia, while the current result from an auditory oddball task shows a decrease in motor activity. The change in direction is likely due to the significant attentional component in the auditory oddball task. In contrast, the sensorimotor task is predictable, and increase in motor activity in patients performing similar tasks, have been noticed in neuroimaging literature.

We also perform CCA-based fusion on the fMRI and sMRI data sets while excluding the EEG data. Comparing the results of the three-way analysis with those from the two-way analysis (results not shown), we find that for both experiments the areas detected in the fMRI and sMRI component are very similar for the component that showed significant differences between the two groups, with the two-way analysis showing some areas of deactivation. Additionally, we note that the statistical significance of the difference between healthy controls and patients increased significantly with the use of three modalities in the analysis as compared to two modalities (Table 1) confirming the expectation that increased number of modalities do help identify more discriminative features increasing the overall sensitivity of the analysis. Also, if the tests are corrected for Type-I errors using the Bonferroni correction, the significance threshold would be 0.003 and we can see in Table 1 that the results of the three modality fusion satisfy this threshold for at least two modalities while the two modality fusion results do not pass the threshold. Also, note that the Bonferroni correction may be too conservative and instead a less conservative false discovery rate threshold can also be used to check the significance of the results.

In the previous sections, we have presented the use of CCA and M-CCA for the fusion of data from different modalities. Next, we present a related but different framework for the use of M-CCA to perform group study of data from the same modality.

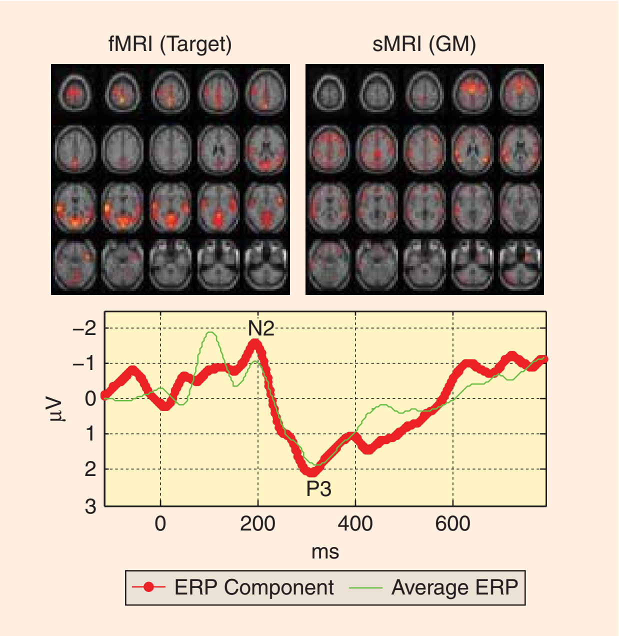
MULTISUBJECT DATA ANALYSIS

In biomedical applications, it is common to study data from a number of subjects under identical experimental conditions and to make inferences based on group analysis—simply looking for occurrences that can be said to be true for the group. In this section, we present the M-CCA based group analysis method for multisubject fMRI data analysis introduced in [19] and highlight one of the key results from the paper along with a comparison with the group ICA analysis technique.

GENERATIVE MODEL AND M-CCA FOR GROUP ANALYSIS

For a group of K data sets, each data set $\mathbf{X}_k = [\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \ldots, \mathbf{x}_k^{(N)}]^T$, $k = 1, 2, \ldots, K$ contains linear mixtures of N sources given in the source vector $\mathbf{S}_k = [\mathbf{s}_k^{(1)}, \mathbf{s}_k^{(2)}, \ldots, \mathbf{s}_k^{(N)}]^T$, mixed by a nonsingular matrix, $\mathbf{A}_k$, i.e.,

$$\mathbf{X}_k = \mathbf{A}_k \mathbf{S}_k, \tag{5}$$



[FIG6] Set of associated components estimated by M-CCA that showed significantly different loading for patients versus controls.

where $\mathbf{X}_k, \mathbf{S}_k \in \mathbb{R}^{N \times Q}$ form the mixture data set and source data set respectively, $\mathbf{A}_k \in \mathbb{R}^{N \times N}$ is a nonsingular square matrix. Note that the mixture data set for multisubject BSS is the whole multidimensional data set and is different from the feature data sets used in the previous sections for data fusion. Sources are uncorrelated within each data set and have zero mean and unit variance, i.e., $E\{\mathbf{S}_k\} = 0$, $k = 1, 2, \ldots, K$ and $E\{S_k S_k^T\} = \mathbf{I}$, $k = 1, 2, \ldots, K$ where $\mathbf{I}$ is the identity matrix. Sources from any pair of data sets $k \neq l; k, l \in \{1, 2, \ldots, K\}$ have nonzero correlation only on their corresponding indices. Without loss of generality, we assume that the magnitude of correlation between corresponding sources are in nondecreasing order, i.e., $r_{k,l}^{(1)} \geq r_{k,l}^{(2)}, \ldots, \geq r_{k,l}^{(N)}$, where $r_{k,l}^{(i)} = E\{\mathbf{s}_k^{(i)}(\mathbf{s}_l^{(i)})^T\}$.

This assumed correlation pattern for latent sources in the generative model can be effectively used to construct a multisubject separation scheme using M-CCA , which is shown in Figure 1(b). In this scheme, the group of sources that have the maximal between-set correlation values are first extracted from the data sets. By removing the estimated sources from the data sets and repeating the correlation maximization

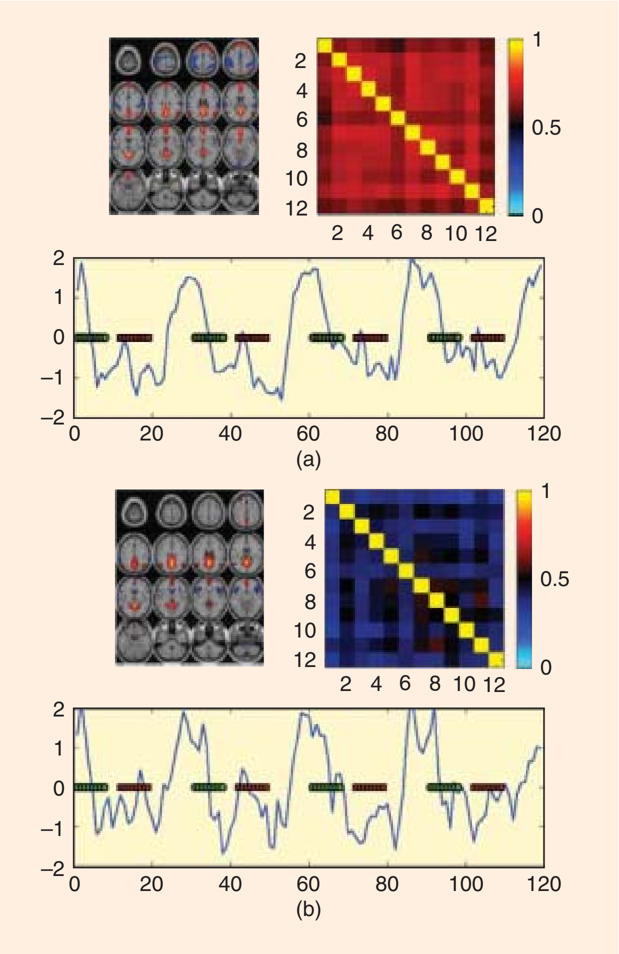
[TABLE 1] COMPARISON OF *t*-TESTS FOR THREE MODALITY (fMRI, sMRI, AND EEG) VERSUS TWO MODALITY (fMRI AND sMRI) ANALYSES FOR COMPONENT. THE *t*-TESTS ARE PERFORMED ON THE LOADINGS FROM THE MODULATION PROFILES OF HEALTHY AND SCHIZOPHRENIC SUBJECTS.

MODALITIES	THREE MODALITIES		TWO MODALITIES	
	t	α	t	α
fMRI	3.45	0.002	2.17	0.038
sMRI	2.86	0.001	2.20	0.034
EEG	3.61	0.007	–	–

procedure, subsequent procedures can extract groups of corresponding sources from each data set in decreasing order of between-set correlation values. This procedure is described in the section "M-CCA," however, in this case the canonical variates will not be defined as in (3) and instead they will be defined as $\mathbf{S}_k = \mathbf{W}_k\mathbf{X}_k$, for $k = 1, 2, \ldots, K$ and $r_{k,l}^{(i)} = E\{\mathbf{s}_k^{(i)}(\mathbf{s}_l^{(i)})^T\}$.

In [19], the study of source separability conditions based on a flexible generative model shows that the method can be used to achieve successful source separation under mild conditions. The conditions depend on the chosen cost function, e.g., we show that $J(r_{k,l}) = \sum_{k,l=1}^{K} |r_{k,l}|^2$ is a practical choice for the cost and that it leads to robust separation performance and a separability condition that is easily satisfied especially when the number of observations increases. The superior performance of M-CCA is shown for group source separation for large number of data sets, robustness to outliers, and robustness to complex-valued data distributions, when compared with data-driven methods that assume a non-Gaussian model [19].

[THE TENDENCY IN DATA ANALYSIS IS TO TRY TO MINIMIZE THE ASSUMPTIONS ON THE NATURE OF DATA, CERTAIN ASSUMPTIONS MAY BE SUITED TO THE DATA BEING STUDIED.]



[FIG7] Estimated mean activation maps (top left), source correlation between subjects (top), and time course (bottom) of the default mode by (a) M-CCA and (b) Group ICA. The right (green circle) and left (red block) visuomotor task paradigm is overlaid onto the estimated time courses for reference.

COMPARISON WITH OTHER APPROACHES

Group ICA achieves source separation of multiple data sets by first reducing the dimensionality of data from each subject, followed by reducing the these reduced data sets to a common subspace, then performing ICA on this common subspace, and finally back-reconstructing the subject-specific source estimates. The data reduction steps used to obtain the common subspace reduces the amount of subject variability in the estimated subject-specific source estimates. M-CCA, on the other hand, performs BSS after a subject-level data reduction stage and does not work on a common subspace. This allows M-CCA to retain much of the intersubject variability. Tensorial ICA also specifies a common signal subspace model that is similar to that of group ICA. Additionally, in tensorial ICA, each group of the corresponding mixing vectors is represented by a common mixing vector associated with a cross-subject variation vector using a rank-one approximation. In this way, the group data sets are decomposed into a three-way tensor product of the common sources, common mixing vectors, and the associated cross-subject variation vectors. IVA uses a mutual information-based formulation to perform source separation across multiple data sets; however, the algorithmic development of the method involves the simplifying assumption that the estimated sources are uncorrelated across data sets, which is an unrealistic assumption for many applications including analysis of biomedical data sets. For a more detailed comparison of these group analysis techniques, refer to [19], and for review of ICA-based multisubject analysis methods in fMRI, refer to [15].

APPLICATION OF M-CCA TO MULTISUBJECT DATA ANALYSIS

The multiple data set extension of CCA, M-CCA reveals relationship among the hidden factors in multiple data sets. In this section, we show how M-CCA can be used for source separation across multiple data sets.

MULTISUBJECT ANALYSIS OF fMRI

Twelve right-handed participants with normal vision—six females, six males, average age 30 years—participated in the study. Subjects performed a visuomotor task involving two identical but spatially offset, periodic, visual stimulus, shifted by 20 s from one another. A total of 12 data sets are jointly analyzed. Each data set is preprocessed according to typical fMRI analysis procedures consisting of slice-timing correction, image registration, motion correction, smoothing, whitening, and dimension reduction. Thirty-two normalized principal components (PCs) are retained for each data set and M-CCA is applied to the 12 sets of retained PCs. The optimal number of PCs are selected using an information theoretic criterion with correction for sample dependence; for implementation details refer to [19].

We present a source of interest from the M-CCA and group ICA estimation results. The M-CCA result shows activation at inferior parietal lobule, posterior cingulate, and medial frontal gyrus—this set of regions is called the “default mode” network that tends to be less active during the performance of a task [34]—as well as deactivation in motor, temporal, and visual regions. The group ICA result, on the other hand, focuses on the default mode activity. The estimated mean activation maps over all data sets, image of the cross-subject source correlation matrices, and the mean time course are displayed in Figure 7. The right- and left-side visuomotor task paradigm is overlaid onto the estimated time courses for reference.

The estimated sources by M-CCA and group ICA are shown in Figure 7. It is observed that the spatial map estimated by M-CCA shows higher cross-subject correlation level than group ICA. The time courses of default mode estimated by M-CCA and group ICA both show expected negative correlation against the onset of the visuomotor task. Furthermore, a multiple linear regression is performed on the estimated time course with the right (R) and left (L) visuomotor paradigm regressors. It is observed that time course estimated by M-CCA has more significant regression coefficients with the task paradigms, i.e., the values are for M-CCA (R): –0.52 with estimated confidence interval (CI): $[-0.38, -0.65]$ and (L): -0.87 CI: $[-0.74, -1.01]$; group ICA (R): -0.45 CI: $[-0.28, -0.62]$ and (L) -0.60 CI: $[-0.43, -0.77]$. Hence, M-CCA achieves higher consistency on spatial activation region and also the time courses show a higher correlation with the task paradigm. The agreement of the spatial and temporal features suggests that default mode network is a common feature across all subjects that is driven by both the left and right visuomotor task.

DISCUSSION

We have presented two CCA-based approaches for data fusion and group analysis of biomedical imaging data and demonstrated their utility on fMRI, sMRI, and EEG data. The results show that CCA and M-CCA are powerful tools that naturally allow the analysis of multiple data sets. The data fusion and group analysis methods presented are completely data driven, and use simple linear mixing models to decompose the data into their latent components. Since CCA and M-CCA are based on second-order statistics they provide a relatively less constrained solution as compared to methods based on higher-order statistics such as ICA. While this can be advantageous, the flexibility also tends to lead to solutions that are less sparse than those obtained using assumptions of non-Gaussianity—in particular super-Gaussianity—at times making the results more difficult to interpret. Thus, it is important to note that both approaches provide complementary perspectives, and hence it is beneficial to study the data using different analysis techniques.

[CCA- AND M-CCA-BASED METHODS PROVIDE ATTRACTIVE SOLUTIONS TO DATA FUSION AND GROUP ANALYSIS.]

Though, in general, the tendency in data analysis is to try to minimize the assumptions on the nature of data, certain assumptions may be suited to the data being studied, and strong assumptions such as independence or sparsity might help improve robustness of the solutions. Thus, the performance of a method should be judged on the overall properties rather than a simple optimality criterion, while taking the underlying assumptions into account. Especially in the case of the study of brain structure and function, since the ground truth is seldom available, the assumptions in most cases cannot be verified. Thus, fully exploiting the complementary nature of different methods becomes especially important, as now noted in most neuroimaging literature. As we demonstrate in this article, CCA- and M-CCA-based methods provide attractive solutions to data fusion and group analysis, and their true power might be realized when they are used in conjunction with other methods that are complementary in nature. Also, their extensions to incorporate sparsity and higher-order statistical information can help improve their utility.

AUTHORS

Nicolle M. Correa (nicolle1@umbc.edu) received a bachelor’s degree in electronics and telecommunications engineering from St. Francis Institute of Technology, Mumbai University, India, in 2003. She received a master’s degree in electrical engineering from the University of Maryland Baltimore County (UMBC), in 2005. She is currently a Ph.D. candidate in the Machine Learning for Signal Processing Laboratory at UMBC. Her research interests include statistical signal processing, machine learning, and their applications to medical image analysis, genetics, and proteomics.

Tülay Adalı (adali@umbc.edu) received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, in 1992 and joined the faculty at UMBC as a professor that same year. She was the general cochair of the IEEE International Workshop on Neural Networks for Signal Processing (2001–2003); technical chair of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2004–2008); publicity chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2000 and 2005); and publications cochair of ICASSP 2008. She chaired the IEEE Signal

Processing Society (SPS) MLSP Technical Committee (2003–2005) and serves on the IEEE SPS MLSP and the Signal Processing Theory and Methods Technical Committees. She was an associate editor for both *IEEE Transactions on Signal Processing* and *Signal Processing* and is currently an associate editor for *IEEE Transactions on Biomedical Engineering, IEEE Journal of Selected Topics in Signal Processing*, and *Journal of Signal Processing Systems*. She is a Fellow of the IEEE and the AIMBE and a recipient of an NSF CAREER Award. Her research interests are in the areas of statistical signal processing, MLSP, and biomedical data analysis.

Yi-Ou Li (liyiou1@umbc.edu) received his B.S. degree in electrical engineering from Beijing University of Posts and Telecommunications. He graduated from UMBC with a Ph.D. degree in electrical engineering. He joined the Neural Connectivity Lab in the Department of Radiology and Biomedical Imaging at the University of California, San Francisco, in 2009. His research interests include statistical signal processing, multivariate analysis, and quantitative evaluation of data-driven methods and their applications to functional MRI data analysis.

Vince D. Calhoun (vcalhoun@unm.edu) received his Ph.D. degree in electrical engineering from UMBC in 2002. He is the chief technology officer at the Mind Research Network and is an associate professor at the University of New Mexico. He is the author of over 100 journal articles and 200 conference proceedings. He is on the editorial board of *Human Brain Mapping* and *Neuroimage*. His research interests include the development of data-driven methods for brain imaging and genetics in the areas of image processing and pattern recognition. He is a Senior Member of the IEEE.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3–4, pp. 321–377, 1936.

[2] J. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[3] K. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Hum. Brain Mapp.*, vol. 1, no. 2, pp. 153–171, 1994.

[4] M. J. Mckeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, and T. J. Sejnowski, "Analysis of fMRI by blind separation into independent spatial components," *Hum. Brain Mapp.*, vol. 6, pp. 160–188, 1998.

[5] V. D. Calhoun and T. Adalı, "Unmixing fMRI with independent component analysis," *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 2, pp. 79–90, Mar./Apr. 2006.

[6] T. P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Analysis and visualization of single-trial event-related potentials," *Hum. Brain Mapp.*, vol. 14, pp. 166–185, 2001.

[7] L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Phys. Rev. Lett.*, vol. 72, no. 23, pp. 3634–3637, 1994.

[8] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Exploratory fMRI analysis by autocorrelation maximization," *Neuroimage*, vol. 16, no. 2, pp. 454–464, 2002.

[9] A. S. Lukic, M. N. Wernick, L. K. Hansen, J. Anderson, and S. C. Strother, "A spatially robust ICA algorithm for multiple fMRI data sets," in *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI)*, 2002, pp. 839–842.

[10] F. Savopol and C. Armenakis,"Mergine of heterogeneous data for emergency mapping: Data integration or data fusion?," in *Proc. ISPRS*, Ottawa, ON, Canada, 2002.

[11] V. D. Calhoun, T. Adalı, G. D. Pearlson, and K. A. Kiehl, "Neuronal chronometry of target detection: Fusion of hemodynamic and event-related potential data," *Neuroimage*, vol. 30, no. 2, pp. 544–553, 2006.

[12] N. Correa, Y.-O. Li, T. Adalı, and V. D. Calhoun, "Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia," *IEEE J. Select. Topics Signal Process.*, vol. 2, no. 6, pp. 998–1007, Dec. 2008.

[13] J. Liu, G. D. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, and V. D. Calhoun, "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Hum. Brain Mapp.*, vol. 30, no. 1, pp. 241–255, 2009.

[14] V. D. Calhoun and T. Adalı, "Feature-based fusion of medical imaging data," *IEEE Trans. Inform. Technol. Biomed.*, vol. 13, no. 5, pp. 711–720, 2009.

[15] V. D. Calhoun, J. Sui, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *Neuroimage*, vol. 45, no. 1 (Suppl. 1), pp. S163–S172, Mar. 2009.

[16] N. Correa, Y.-O. Li, T. Adalı, and V. D. Calhoun, "Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 385–388.

[17] A. A. Nielsen, "Multiset canonical correlation analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, 2002.

[18] Q.-S. Sun, S.-G. Zeng, P.-A. Heng, and D.-S. Xia, "Feature fusion method based on canonical correlation analysis and handwritten character recognition," in *Proc. Int. Conf. Control, Automation, Robotics, and Vision*, 2004, vol. 2, pp. 1547–1552.

[19] Y.-O. Li, W. Wang, T. Adalı, and V. D. Calhoun, "Joint blind source separation by multi-set canonical correlation analysis," *IEEE Trans. Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[20] V. D. Calhoun, T. Adalı, J. J. Pekar, and G. D. Pearlson, "A method for making group inferences from functional MRI data using independent component analysis," *Hum. Brain Mapp.*, vol. 14, no. 3, pp. 140–151, 2001.

[21] C. F. Beckmann and S. M. Smith, "Tensorial extensions of independent component analysis for group fMRI data analysis," *Neuroimage*, vol. 25, no. 1, pp. 294–311, 2005.

[22] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[23] M. S. Cohen, R. M. DuBois, and M. M. Zeineh, "Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging," *Hum. Brain Mapp.*, vol. 10, no. 4, pp. 204–211, 2000.

[24] G. Gerig, J. Martin, R. Kikinis, O. Kubler, M. E. Shenton, and F. Jolesz, "Unsupervised tissue type segmentation of 3D dual-echo MR head data," *Image Vis. Comput.*, vol. 10, no. 6, pp. 349–360, 1992.

[25] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[26] W. M. Wells, III, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imaging*, vol. 15, no. 4, pp. 429–442, 1996.

[27] J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition. [Review]," *Med. Phys.*, vol. 20, no. 4, pp. 1033–1048, 1993.

[28] V. D. Calhoun, T. Adalı, N. R. Giuliani, J. J. Pekar, K. A. Kiehl, and G. D. Pearlson, "Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data," *Neuroimage*, vol. 27, pp. 47–62, 2006.

[29] G. Machado, M. Juarez, V. P. Clark, R. Gollub, V. Magnotta, T. White, and V. D. Calhoun, "Probing schizophrenia using a sensorimotor task: Large-scale (N=273) independent component analysis of first episode and chronic schizophrenia patients," in *Proc. Society for Neuroscience*, San Diego, CA, 2007.

[30] G. D. Pearlson and L. Marsh, "Structural brain imaging in schizophrenia: A selective review," *Biol. Psychiatry*, vol. 46, no. 5, pp. 627–649, 1999.

[31] G. D. Pearlson and V. D. Calhoun, "Structural and functional magnetic resonance imaging in psychiatric disorders," *Can. J. Psychiatry*, vol. 52, no. 3, pp. 158–166, 2007.

[32] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1004–1034, 1995.

[33] K. A. Kiehl, M. Stevens, K. R. Laurens, G. D. Pearlson, V. D. Calhoun, and P. F. Liddle, "An adaptive reflexive processing model of neurocognitive function: Supporting evidence from a large scale (n=100) fMRI study of an auditory oddball task," *Neuroimage*, vol. 25, no. 3, pp. 899–915, 2005.

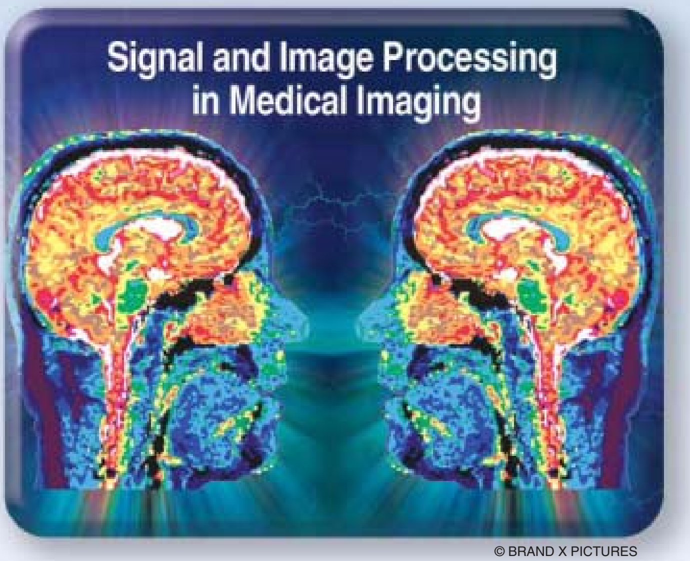
[34] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function," *Proc. Nat. Acad. Sci.*, vol. 98, no. 2, pp. 676–682, 2001.

[SP]

[Dzung L. Pham, Pierre-Louis Bazin, and Jerry L. Prince]

Digital Topology in Brain Imaging

[Creating anatomically consistent representations of the human brain]



© BRAND X PICTURES

Modeling topology in medical image processing algorithms has emerged as a powerful technique for computing structural representations that are consistent with the underlying anatomy. When applied to high resolution images of the brain, these methods have proven to be extremely beneficial to neuroscientific studies in generating mathematical representations of the cerebral cortex and other brain structures, improving the analysis and visualization of functional activity, and allowing for group comparisons of brain geometry. Topological properties help model the global connectivity of structures without placing a bias on shape. In addition to providing anatomical consistency, topology-preserving algorithms also exhibit an improved robustness to noise. We provide an introduction to the main concepts in digital topology on which these algorithms are based and review their use in the segmentation of magnetic resonance (MR) brain images.

BACKGROUND AND SIGNIFICANCE

With the advent of high resolution MR imaging technologies, the development of algorithms to segment and reconstruct the human brain has been an important area of interest for both the medical image processing and neuroscientific communities. In particular, the reconstruction of the cerebral cortex from MR images allows neuroscientific researchers to better understand the structure and function of the brain in health and disease [1], [2]. The cerebral cortex is a highly convoluted layer of gray matter tissue within the human brain that is known to be responsible for motor, sensation, and cognitive processing. Because the cortex is a very thin structure with complex folding patterns, accurate reconstruction is a challenging problem. Furthermore, the ability of any cortical representation to be "flattened" or "unfolded" is paramount to both visualization of functional activity mapped to the cortex, as well as generating a standardized space for performing group comparisons [3]. Such a representation must possess a topology that is consistent with the known anatomy.

The inner and outer cortices, when connected across the brain stem, are known to each be topologically equivalent to a sphere [3]. This is illustrated in Figure 1(a), where the left side shows a coronal section of an MR brain image with the inner and outer cortices highlighted in yellow and the right side shows a cross section of two concentric spheres. This topological equivalence implies that any cortical surface may be

Digital Object Identifier 10.1109/MSP.2010.936729

1053-5888/10/$26.00©2010IEEE

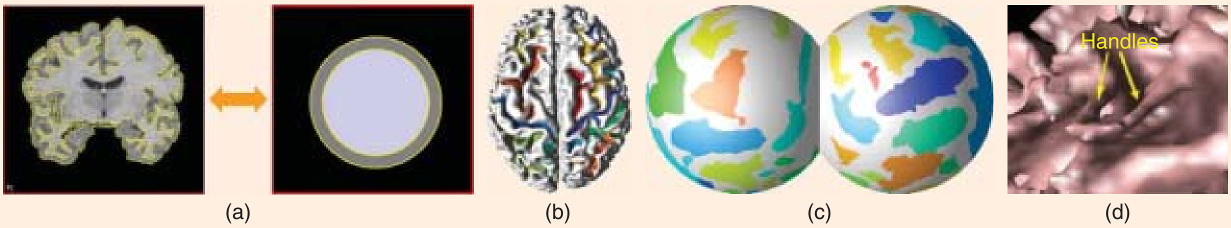
deformed into a sphere without the need to remove or add disconnected pieces, or make tears within the surface. Figure 1(b) shows a topologically correct reconstruction of the cerebral cortex where sulcal regions have been labeled with different colors. Such a reconstruction can be unfolded, as shown in Figure 1(c), exposing the buried sulcal regions and allowing for better visualization of functional activity. Unfolded cortical surfaces may also be used to form a standardized coordinate system for group analyses [3]. Reconstruction approaches that ignore the issue of topology may result in cortices composed of multiple pieces or surfaces that possess handles, which are known to be anatomically invalid and are shown in Figure 1(d).

Three approaches are possible for obtaining segmentations of a single object with a desired topology: 1) topology correction approaches, 2) topology-preserving segmentation approaches, and 3) hybrid approaches that combine both topology correction and topology-preserving segmentation. Topology correction generates a cortical reconstruction using standard methods without consideration of topology, and then applies a correction algorithm. Early work in performing topology correction of cortical reconstructions involved laborious amounts of manual editing to ensure that the cortical surface was a single connected piece with no handles [3], [1]. This eventually led to the development of a number of automated topology correction methods (see the section "Topology Correction"). Topology constraints may also be incorporated directly into the segmentation algorithm. To obtain the desired topology, the algorithm is initialized from a template object with the correct topology and then applies topology-preserving deformations. Parametric deformable surface models were commonly used for this purpose because they are inherently topology-preserving, but topology-preserving level set and region growing methods have also been proposed (see the sections "Voxel-Based Methods" and "Deformable Models"). These methods may suffer, however, from sensitivity to initialization and convergence to local optima, leading into inaccurate results. Hybrid approaches addressed this sensitivity by combining topology correction with topology-preserving deformable models. For cortical surface reconstructions, the most common approach utilizes a topologically corrected representation of the white matter surface that is then used as an initialization for a topology-preserving deformable model [1], [2].

Enforcing topology preservation or correction is nontrivial. Since topology is known to be a global property, it can potentially be extremely computationally expensive to monitor. By employing fundamental concepts from the field of digital topology, topological changes can be efficiently evaluated using only local computations [4] (see the section "Digital Topology"). Thus, although topology-preserving algorithms do require some additional computational overhead when compared to its traditional counterpart, the additional expense is typically small.

Although reconstruction of the cerebral cortex was the original motivation for many of the image processing methods involving topological models, nearly all anatomical structures in the healthy brain and the human body can be assumed to have a fixed topology. Standard brain tissue segmentation techniques can often result in a single pixel labeled as gray matter in the middle of white matter even though it is anatomically known that such configurations cannot exist. Local smoothing priors such as Markov random fields [5] can penalize against such configurations but can not completely eliminate them without sacrificing accuracy due to oversmoothing. Topological modeling, on the other hand, has the ability to explicitly prevent such configurations from occurring. Validation experiments have shown that algorithms incorporating topological constraints have increased robustness to noise over standard approaches without loss of accuracy [6], [7].

Topology-preserving techniques have had a tremendous impact on the neuroimaging and neuroscientific communities by enabling quantitative measures to be derived, particularly from the cortex. These measures, such as cortical thickness, have been shown to be useful in characterizing aging processes and diseases (cf. [8]). Research in this area has also spawned new theoretical developments in both digital topology and mathematics [9]. The Web site of the cortical reconstruction software package FreeSurfer [1] alone lists over 70 methodological and neuroscientific publications based on its tools. In the following, we introduce the reader to the basic digital topology concepts utilized in the development and implementation of topologically constrained image processing algorithms. An overview of these methods and their application to human brain mapping is then provided. We focus primarily on three-dimensional (3-D) cases, although some two-dimensional (2-D) examples are provided for ease of visualization.



[FIG1] Topology in human brain mapping: (a) A 2-D view illustrating topological equivalence of the cerebral cortex and the sphere, (b) reconstructed cortical surface with sulcal regions assigned to color labels, (c) unfolded cortical surface with each hemisphere mapped to a spherical coordinate system, and (d) magnified view of a cortical surface without topology constraints.

BASIC CONCEPTS AND METHODS

In this section, we define basic terminology and introduce important concepts for constraining topology in image processing methods. A critical consequence of results from digital topology is that changes in topology can be detected using local computations.

TOPOLOGY IN CONTINUOUS DOMAINS

Topology is a vast domain of mathematics, classically considered to be founded by Euler in 1736 with his solution to the Konigsberg bridge problem [10]. Since then, the subject has permeated many aspects of applied and abstract mathematical theory. Topology is defined as the spatial property retained by an object under any continuous geometric transformation, including bending, twisting, stretching, but not including tearing or joining. When considering a 3-D object bounded by a closed surface, the surface can move in any way that does not create a cut or a self-intersection in the surface. This means, for instance, that the cortical surface of any healthy brain can be deformed into a sphere or into any other cortical surface without altering topology.

The topology of a set of 3-D objects is fully characterized by the number of disjoint closed simple surfaces c, representing the boundaries of the objects, as well as the number of closed loops or handles in each object, called the genus g of a surface. These two numbers can be used to formulate the Euler characteristic χ. A more practical equation for the Euler characteristic can also be computed for any mesh parameterization of the surfaces as a function of the number of vertices V, edges E, and faces F [11], [12]

$$\chi = 2c - 2g = V - E + F. \tag{1}$$

Spherical topology refers to the topology of the sphere, which is the simplest possible case where $c = 1$, $g = 0$, and $\chi = 2$. Figure 2 shows an example of a torus, where $c = 1$, $g = 1$, and $\chi = 0$. The torus can be digitally represented as a surface mesh [Figure 2(b)]. The same value of zero would be derived regardless of whether the mesh shown in Figure 2 was refined, or if an alternative tessellation had been used, such as a triangular or simplex mesh. The torus can also be represented as a binary 3-D image, shown in Figure 2(c). An important difference between these two digital representations is that the vertices of the surface may exist in a continuous space, while the pixels of the image exist on a discrete grid.

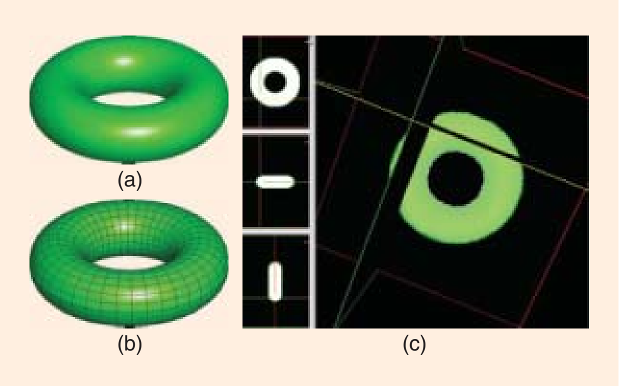
Topology is a global property in that the overall topology of an object can not be determined from only a portion of the object. However, it is possible to determine whether the topology of an object has changed based on local computations. A geometric transformation applied to an object will preserve the object's topology if and only if the transformation is a homeomorphism, a continuous function with a continuous inverse defined on the original and deformed objects. In Euclidean spaces up to dimension three, homeomorphisms are also diffeomorphisms, invertible functions for which both the function and its inverse are differentiable. Diffeomorphisms are more convenient to satisfy mathematically in constructing a topology-preserving transformation since they may be enforced by ensuring that the Jacobian of the transformation exists and is strictly nonzero everywhere within the object space [13].

DIGITAL TOPOLOGY

Although the definitions above apply to surface representations and transformations within the continuous domain, their extension to the discretely sampled spaces of digital images is nontrivial. A continuously diffeomorphic transformation may not preserve topology in the digital space. For example, depending on connectivity and sampling assumptions, a simple rotation or scaling could cause a thin digital object to break into multiple parts [14].

Digital topology, which was pioneered by Azriel Rosenfeld in the latter half of the 20th century, bridges the gap between continuous and digital spaces and offers fundamental tools for handling the topological properties of digital images [4]. At the core of image processing methods utilizing digital topology is the notion of a simple point, which is a pixel that can be freely labeled as either inside or outside an object without changing the topology of the object. Consider an object represented by a binary digital image. By definition, any transformation that flips the label of one pixel within the binary image preserves topology if and only if that pixel is a simple point. Figures 3 and 4 show examples of simple and nonsimple points in 2-D and 3-D. Where in the continuous domain diffeomorphic conditions could be used to enforce topology-preserving transformations, these are are replaced by considerations of simple points in the digital domain.

Simple points can be identified in a number of ways by considering a local neighborhood of the point. An elegant and efficient approach involves the notions of geodesic neighborhoods and topological numbers [15]. The geodesic neighborhood $N_n^k(\mathbf{x}, X)$ of a point $\mathbf{x}$ in an object region X is the set of pixels with the same binary label for which there is a n-connected path in X of length no greater than k between the neighbor and $\mathbf{x}$. The neighborhood depends on a choice of connectivity, which determines how pixels are connected inside or outside the object. In 2-D, 4-connectivity implies that only horizontal and vertical



[FIG2] (a) Torus object, (b) torus represented as a rectangular surface mesh, and (c) torus represented as a 3-D binary image volume, with orthogonal cross sections shown.

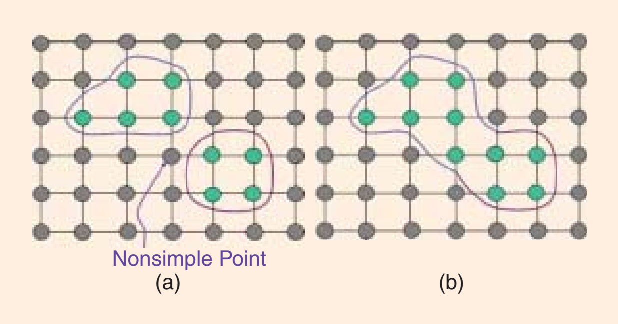
neighbors can be connected, while 8-connectivity includes diagonal neighbors. The 3-D connectivities are shown in Figure 5(a), where the red circles indicate six-connectivity, the red and blue circles combined form 18-connectivity, and all circles surrounding the center point form 26-connectivity. Connectivities work in complementary pairs to avoid gaps (regions neither inside nor outside the object) or overlaps (regions both inside and outside the object) in the continuous representation of the object. For example in Figure 3, if the object and background were both 8-connected, they would overlap one another. Connectivity assumptions may be either 4/8 (object/background) or 8/4 in 2-D, and 6/18, 6/26, 18/6, or 26/6 in 3-D.

The topological numbers of x relative to the set X are the numbers of connected components within the geodesic neighborhoods defined by the choice of a connectivity rule. In two dimensions, the topological numbers T_n are given by

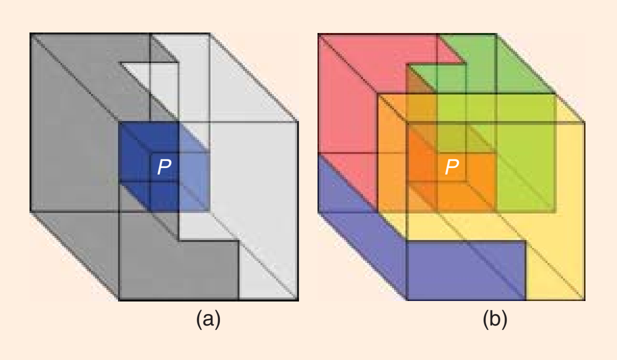
$$T_4(\mathrm{x}, X) = \mathcal{C}_4(N_4^2(\mathrm{x}, X))$$
$$T_8(\mathrm{x}, X) = \mathcal{C}_8(N_8^1(\mathrm{x}, X)),$$

where $\mathcal{C}_n(X)$ denotes the number of n-connected components in X. In three dimensions, we have

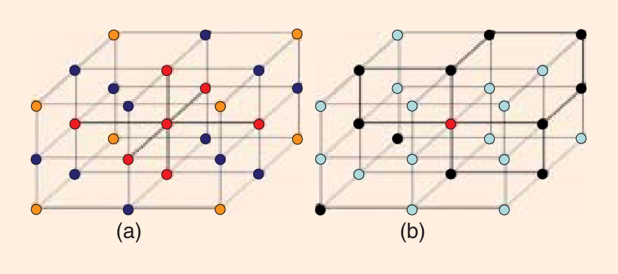
$$T_6(\mathrm{x}, X) = \mathcal{C}_6(N_6^2(\mathrm{x}, X))$$
$$T_{6+}(\mathrm{x}, X) = \mathcal{C}_6(N_6^3(\mathrm{x}, X))$$
$$T_{18}(\mathrm{x}, X) = \mathcal{C}_{18}(N_{18}^2(\mathrm{x}, X))$$
$$T_{26}(\mathrm{x}, X) = \mathcal{C}_{26}(N_{26}^1(\mathrm{x}, X)).$$



[FIG3] Illustration of a nonsimple point: if the nonsimple point changes labels, then the two 4-connected objects (left) merge to form a single 4-connected object (right).



[FIG4] (a) 3-D illustration of a simple point and (b) multiobject generalization of a simple point (see the section "Multiobject Topology").



[FIG5] (a) Connectivity in 3-D and (b) an example of computing topological numbers.

There are two topological numbers in the 3-D case for 6-connectivity, following the convention introduced in [15], wherein the notation "6+" implies 6-connectivity whose dual connectivity is 18, while the notation "6" implies 6-connectivity whose dual connectivity is 26. This highlights the fact that connectivity fundamentally works in pairs, and so the topological number for 6-connectivity must be computed differently depending on whether the associated connectivity rule is 18 or 26.

The topological numbers allow straightforward characterization of a simple point. It is proven in [15] that a point x is simple if and only if $T_n(\mathrm{x}, X) = 1$ and $T_{\bar{n}}(\mathrm{x}, \bar{X}) = 1$, where $(n, \bar{n})$ is a pair of compatible connectivities and $\bar{X}$ represents the background object. In other words, a point is simple if there is exactly one connected inside region and one connected outside region in the neighborhood of the point. A straightforward computation of topological numbers can be implemented by counting the number of connected components within specific geodesic neighborhoods. A more efficient approach is to build a look-up table to store the topological numbers for each possible configuration. This approach can be memory intensive for the 3-D case since the table could have 2^{26} entries. Another approach based on binary decision diagrams avoids the use of large look-up tables by performing an efficient series of tests [16]. In Figure 3, $T_4 = 2$ for the object and $T_{\bar{8}} = 2$ for the background, implying the point is not simple. However, under the reverse connectivity assumption, $T_8 = T_{\bar{4}} = 1$, and the point is simple. In Figure 5(b), where black circles are the object and blue circles are the background, $T_6 = 2$ and $T_{\overline{26}} = 1$ so the point is not simple under 6/26 connectivity, but for 6/18 connectivity, $T_{6+} = T_{\overline{18}} = 1$ and the point is simple.

TOPOLOGY-PRESERVING ALGORITHMS

In this section, we provide an overview of topology-preserving models, topology correction, and other recent approaches in the topologically constrained processing of brain images.

VOXEL-BASED METHODS

The concept of the simple point immediately lends itself to voxel-based segmentation approaches involving a region growing or similar process that begins with an initial structure with the desired topology, and expands that structure in a topology-preserving fashion until it segments the desired anatomy. Mangin et al. were among the first to apply such an approach in segmenting multiple structures within brain images [17]. Their homotopic (i.e., topology-preserving) deformable region utilized a combination of regularization based on a Gibbs energy function and conditional morphological filters to extract the union of gray matter and cerebrospinal fluid in the brain, which was assumed to have the topology of a sphere with a single cavity. A homotopic skeletonization was then applied to extract sulcal regions within the cerebral cortex. In [6], topology-preserving fast marching algorithms were implemented in combination with an intensity-based clustering algorithm to segment multiple regions of the brain with any desired topology. The topology of each region was dictated by a topology template, which provides a simple representation of the brain structure. Vascular segmentation in both the brain and other anatomical regions have also been performed using topology-preserving region growing approaches [18] under the assumption that vessels are simply connected, and do not possess loops or cavities. In [19], a graph cuts segmentation algorithm was proposed that employed a topology-preserving prior. This formulation takes advantage of the well-known computational efficiency and convergence properties of discrete graph-based segmentation algorithms.

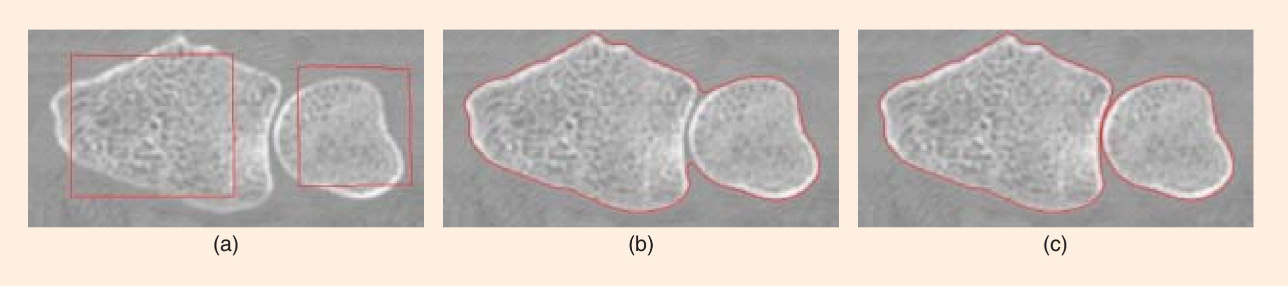
DEFORMABLE MODELS

Deformable models are object-delineating curves or surfaces that move within 2-D or 3-D digital images under the influence of both internal and external forces and user defined constraints [20]. These algorithms have been at the heart of one of the most active and successful research areas in edge detection, image segmentation, shape modeling, and visual tracking. Deformable models are broadly classified as either parametric deformable models or geometric deformable models according to their representation and implementation. In particular, parametric deformable models are represented explicitly as parameterized curved or surfaces in a Lagrangian framework. Geometric deformable models, on the other hand, are represented implicitly as level sets of higher-dimensional, scalar functions and evolve in an Eulerian fashion.

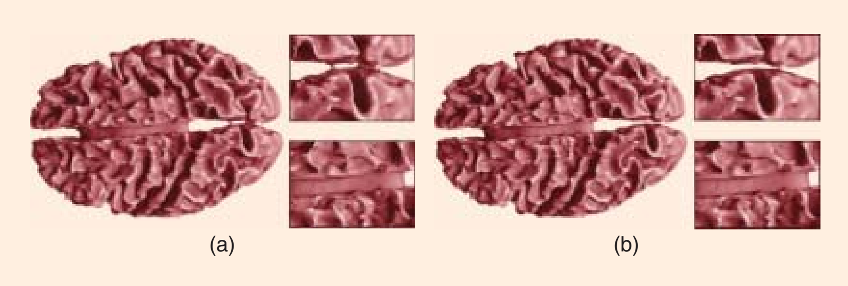
Parametric deformable models are inherently topology preserving because of their Lagrangian formulation. They are typically initialized as closed simple curves or surfaces and remain as such throughout their deformation. An advantage of these models is that they exist in the continuous domain and possess subvoxel accuracy. For these reasons, parametric deformable models have been used extensively in the reconstruction of the cortical surface [21], [1]. However, special care must be paid to preventing self-intersections during the evolution of the model. In [21], self-intersections were prevented using a self-proximity term within the energy function that assigned an increasingly higher cost as the faces of the surface approached each other.

Geometric deformable models have several important advantages over parametric models. First, they are completely intrinsic and therefore are independent of the parameterization of the evolving contour. In fact, the model is generally not parameterized until evolution of the level set function is complete. Thus, there is no need to add or remove nodes from an initial parameterization or adjust the spacing of the nodes as in parametric models. Because parameterization is not required during evolution, self-intersections can easily be prevented. Second, the intrinsic geometric properties of the contour such as the unit normal vector and the curvature can be computed from the level set function in a straightfoward fashion. This contrasts with the parametric case, where inaccuracies in the calculations of normals and curvature result from the discrete nature of the contour parameterization.

Traditional geometric deformable models automatically change topology during evolution. This flexibility is a major advantage of geometric deformable models over parametric deformable models in many applications. To counter this advantage, methods to adaptively change contour topology have also been developed for parametric deformable models [20]. As we have described however, topological flexibility can also be a disadvantage, particular when the anatomy under study is known to be fixed. In [22], topology preservation was achieved by verifying that only simple points were allowed to change sign within the level set at each step of the evolution. This approach maintains the subpixel interpolation and boundary regularization properties of geometric deformable models and is computationally much more efficient than parametric models. Figure 6 shows a 2-D example of segmenting bones using a geometric deformable model with and without topology constraints. Because of the proximity of the bones, the unconstrained model merges to form a single contour, while the topology-preserving approach maintains a clear separation.



[FIG6] Level set segmentation of bone: (a) initialization, (b) standard geometric deformable model, and (c) topology-preserving geometric deformable model.



[FIG7] Isosurfaces computed from a segmentation of the inner cortex (a) before and (b) after topology correction with two magnified views of each to the right. (Images provided courtesy of David Shattuck.)

If a surface representation is required from a geometric deformable model, an isosurface algorithm can be performed, such as the marching cubes technique [23]. The recovered surface is a geometric approximation that will possess the correct topology if the implicit surface is assumed to cross at most once through each edge between two neighboring points, and if the isosurface connectivity is consistent with the choice of digital connectivity on the grid [22]. Although the marching cubes reconstruction performs only a trilinear interpolation of space, more elaborate interpolation techniques may introduce small perturbations on the level set function that become topology artifacts, thereby severing the link between the continuous and digital representations. The other interpolation method that will preserve topological properties is the nearest neighbor interpolation, which represents each sample point inside the object as a voxel.

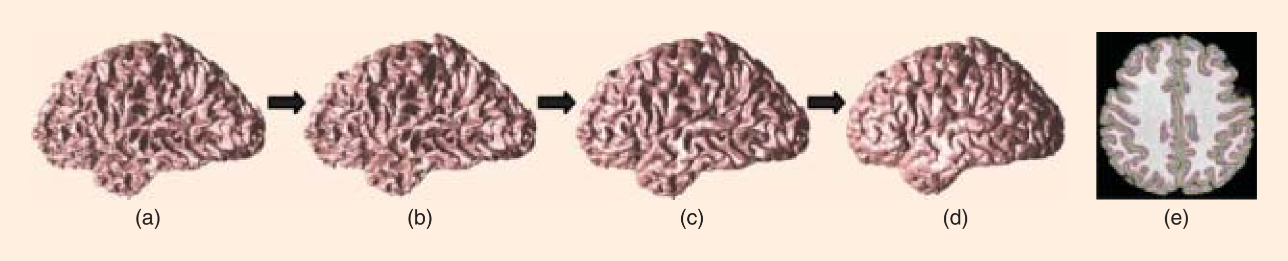
TOPOLOGY CORRECTION

An alternative to topology-preserving segmentation approaches is to apply a segmentation approach that does not consider topology, and then automatically correct the resulting topology. Topology correction methods are often used in combination with topology-preserving segmentation. The latter can suffer from convergence to local optima, particularly because the topology constraints will restrict the freedom of evolving a region or surface. An initialization with the proper topology close to the final structure of interest is therefore quite desirable to improve the accuracy of the resulting delineation. Such hybrid methods are commonplace in the reconstruction of the cerebral cortex [1], [2].

Automated algorithms for topology correction typically operate on a binary volume extracted from the classification of the image data. In these approaches, the objective is to detect and remove topological defects (disconnected pieces and handles) from a segmentation with arbitrary topology, while effecting minimal changes. Most methods first identify the largest connected component of the segmentation, and assume that this component should possess a spherical topology. The manner in which the final topology is enforced can be classified as 1) graph-based analysis [24]–[26], 2) region-growing based [27]–[29], and 3) surface based [30], [31]. An approach also exists for correcting continuously valued object representations that may be obtained in probabilistic or fuzzy segmentation approaches [32]. This method ensures that all isolevels within the segmentation will have the appropriate topology.

Graph-based techniques represent the connected components of an object as graphs that are then processed to identify and remove cycles, which correspond to handles. In [26], background and foreground connectivity graphs were generated by assigning each connected component within a two-dimensional slice to a vertex or node in the graph. Connections in the graph were formed based on whether adjacent slices shared a face within contiguous voxels. The authors conjectured that the object's surface had a spherical topology if both the foreground and background connectivity graphs were both trees. This conjecture was later mathematically proven to be true under mild conditions [33]. Figure 7 shows an example of isosurfaces computed from topologically uncorrected and corrected segmentations of the inner cerebral cortex (i.e., the gray matter/white matter interface).

Region-growing methods start from some initialization, which can be seed points [29], [28] or a bounding box [27], that grows in a topologically controlled fashion. Multiple objects can be simultaneously corrected in this fashion with each object assumed to be homeomorphic to a sphere [28]. Both region-growing and graph-based approaches operate on the digital image and require a subsequent isosurface generation if a surface is desired. On the other hand, surface-based approaches operate directly on the mesh representations. Surface correction can be accomplished by mapping the surface to a sphere, detecting the overlapping faces that occur from topological defects, and then retessellating those regions accordingly [31], [30]. Surface-based approaches are able



[FIG8] Cortical surface reconstruction using a hybrid approach: (a) white matter isosurface after topology correction, (b) inner cortical surface, (c) central cortical surface, (d) outer cortical surface, and (e) crosssections of the latter three surfaces overlaid on the MR brain image.

to utilize information about the geometry of the reconstructed surface at the cost of greater computational expense.

Figure 8 shows the result of a combination of topology correction and topology-preserving techniques to generate a nested surface representation of the cerebral cortex [2]. The brain images were initially segmented using an intensity-based classifier that does not constrain topology. This yields a white matter isosurface that is corrected with a graph-based correction technique, followed by refinements using a topology-preserving deformable model to reconstruct the inner, central, and outer cortical layers.

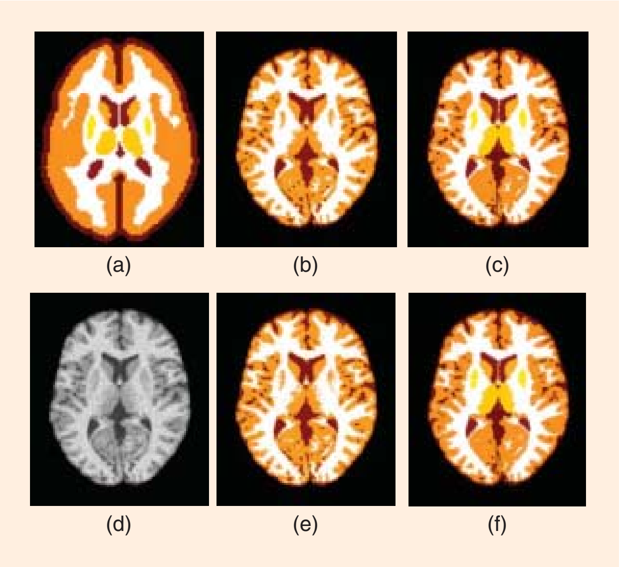
MULTIOBJECT TOPOLOGY

Early work on topologically constrained segmentation of multiple objects focused exclusively on spherical topologies [17]. However, not all objects in the brain possess a spherical topology. White matter, for example, because of its relationship with subcortical gray matter structures and the ventricles, should not be considered to have a spherical topology. An algorithm that performed simultaneous classification of multiple objects with any topology was later proposed [6]. However, these methods only preserved the topology of each object, leaving the topology of the union of any subset of objects to be unconstrained. This is an important issue since the true anatomy typically follows strict topological relationships in terms of how one structure is connected to another. In algorithms that preserve group topologies, neighboring structures in the anatomy must remain neighbors in the segmented image, and nonneighboring structures must stay separated.

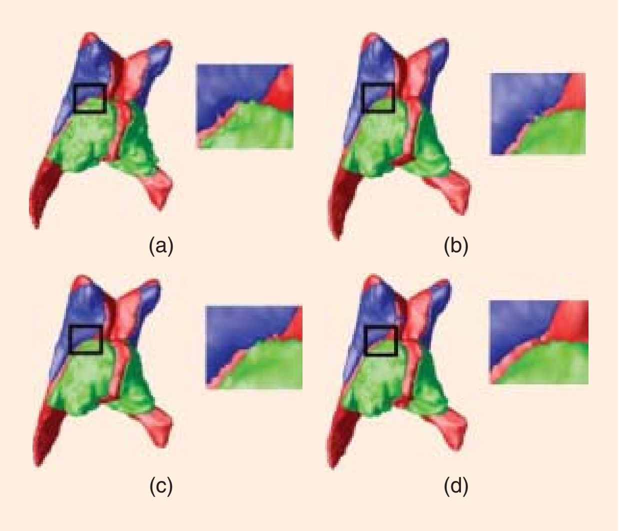
Nonlinear registration techniques (cf. [34]) that volumetrically deform a labeled template image to the subject image and preserve topology using diffeomorphic transformations could be potentially be used for computing topologically constrained segmentations of multiple objects. However, care must be taken in that the diffeomorphism is enforced in a manner consistent with the object representation. In [14] and [35], it is shown that standard diffeomorphic registration techniques will violate the topology of digital objects represented on a pixel or voxel grid, thereby requiring alternative criteria to preserve topology in volumetric deformations.

In [7], topological constraints on both the structures and their groupings were used to encode continuity and relationships without biasing shape, resulting in a strictly homeomorphic segmentation algorithm. The method employed a coarse statistical atlas of shape, and computed the segmentation of cortical and subcortical structures predominantly from the image and topological constraints. To preserve group topologies, a generalization of the simple point, called the digital homeomorphism constraint [14] was employed. This constraint specifies for segmented digital images with multiple labels, the criteria under which pixels may change labels without altering group topology. The constraint can be implemented by performing a simple point check on a limited combination of unions of labels [14]. An illustration of the multiobject case is shown in Figure 4(b).

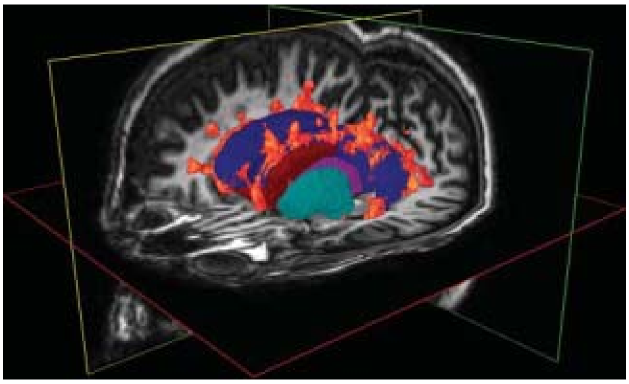
Multiobject segmentation requires a template that describes the desired topology of each structure and how they are connected. In [7], this template was a geometrically simplified representation of the brain and is shown in Figure 9(a). Gray matter is depicted in orange, cerebrospinal fluid in brown, white matter in white, and subcortical gray matter in yellow. One of the benefits of topology constraints is an increased robustness to noise. Figure 9(b)–(f) shows the results of applying an unconstrained segmentation procedure and the multiobject topology-preserving approach when applied to a simulated MR brain image with varying levels of noise. Equivalent amounts of regularization was employed in



[FIG9] Multiobject segmentation: (a) topology template, (b) unconstrained segmentation of simulated image with 3% noise, (c) topology-preserving segmentation of simulated image with 3% noise, (d) simulated image with 7% noise, (e) unconstrained segmentation with 7% noise, and (f) topology-preserving segmentation with 7% noise.



[FIG10] Multiobject segmentation with various constraints: (a) no topology constraints and no smoothing, (b) smoothing only, (c) single object topology constraints, and (d) multiobject topology constraints.



[FIG11] **Example of semiconstrained topology in the segmentation of a multiple sclerosis brain image. Lesions shown in orange may possess any topology but other structures remain constrained.**

each case. Even with increased noise, the topology-preserving approach exhibits no speckling within a structure and only loses accuracy around boundaries.

In [36], a deformable model segmentation framework was described for multiple objects. Using a compact representation of level set functions for multiple objects, the resulting segmentation has no overlap or vacuum, and is based on a computationally efficient evolution scheme independent of the number of objects. Furthermore, the framework can be employed with or without topology control, which is enforced using the digital homeomorphism constraint. Figure 10 shows a 3-D visualization of a segmentation using this method of ventricles (red), caudate (green), and thalamus (blue) from an MR brain image. The caudate and thalamus are known to be in close proximity with one another but do not directly touch. Figure 10(a) has no topology constraint and no local smoothing, Figure 10(b) employs only local smoothing, Figure 10(c) constrains the topology of each individual object, and Figure 10(d) employs multiple object topology constraints. As can be seen, only the latter figure shows a correct reconstruction where the caudate and thalamus do not touch.

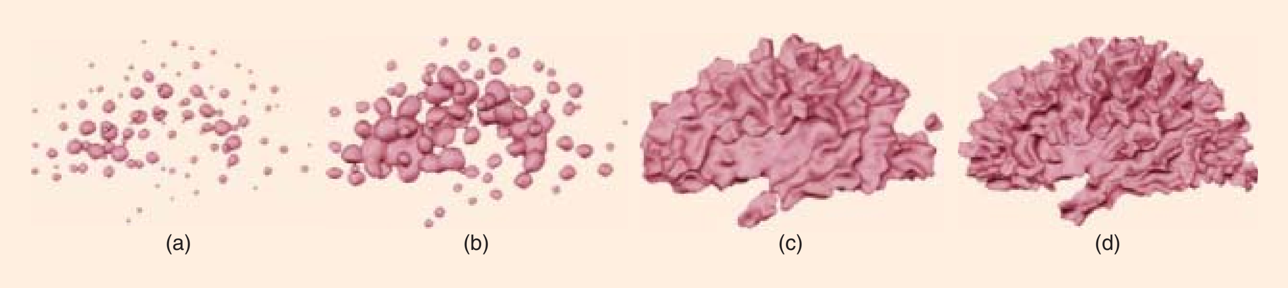
ALTERNATIVE TOPOLOGY CONSTRAINTS

Although topology can be assumed to be fixed in healthy anatomy, disease processes can often alter the topology. In the brain, lesions or tumors may occur that alter the connectivity between different structures, thereby rendering standard topology-preserving methods unsuitable. There have been recent efforts in relaxing topology constraints to account for such situations. In [37], an approach was described for the segmentation of brain images acquired from subjects with multiple sclerosis, which may cause lesions to form within the white matter. These lesions can possess any arbitrary topology. Topology constraints were applied to this problem by assuming that the union of white matter and lesions were fixed, but the lesions themselves were unconstrained. Figure 11 shows an example of this approach.

There has been recent interest in enforcing topology constraints in alternative ways within deformable model approaches. In [38], rather than simply stopping the surface from evolving at the last step before a topology change, which can cause abrupt features to occur in the resulting curve or surface, a topology preserving geometric flow is proposed. This flow imposes a global regularity that allows the deformable model to evolve naturally while preventing topology changes from occurring. Another technique allows geometric deformable models to be split and merge while arriving at a segmentation with approximately the desired topology [39]. Such an approach allows for greater flexibility in the evolution and can potentially prevent convergence to suboptimal configurations. Figure 12 shows an example of this approach in reconstructing the inner cortex.

DISCUSSION

Over the past two decades, great progress has been made in the research of digital topology in brain image analysis. Advances in the incorporation of topology constraints within medical image processing algorithms now allow computationally efficient segmentations that are consistent with the underlying anatomy. Modeling of topology complements commonly used models of local smoothness, such as Markov random fields, and statistical models of shape. We believe topological models will eventually reach similar levels of adoption into image processing algorithms as those other models. Although we have attempted to provide a thorough overview of work in this area, many important contributions have been omitted because of space considerations. A continually updated Web page has been created by the authors that provides a more comprehensive listing of related papers and resources, including publicly available software tools for performing topologically constrained image processing [40].



[FIG12] **(a)–(d) Genus-preserving level set model initialized as a set of 55 spheres and converging to a cortical surface. (Images provided courtesy of Florent Segonne.)**

AUTHORS

Dzung L. Pham (dzung.pham@nih.gov) received the B.S. degree in 1993 from George Washington University and the M.S.E. and Ph.D. degrees in 1995 and 1999, respectively, from Johns Hopkins University, all in electrical engineering. He received an NIH Fellows Award for Research Excellence in 1996 and has coauthored several book chapters in *Handbook of Medical Imaging* from SPIE Press and *Handbook of Medical Image Processing* from Elsevier. He currently directs the Image Processing Core for the Center for Neuroscience and Regenerative Medicine.

Pierre-Louis Bazin (pbazin1@jhmi.edu) received an engineering diploma from Supelec, Orsay, France, in 1998, a D.E.A. from Paris XI University in 1998, and a Ph.D. degree from Paris XI University and INRIA Rocquencourt, France, in 2001. He received post-doctoral training at Brown University from 2001 to 2003 and received an NIH Mentored Quantitative Research Development Award in 2008 for his work, "Topology-Driven Analysis of Brain Anatomy." He is the director of the Laboratory of Medical Image Computing and an instructor in the Department of Radiology and Radiological Science at Johns Hopkins University.

Jerry L. Prince (prince@jhu.edu) received a B.S. degree from the University of Connecticut in 1979 and the S.M., E.E., and Ph.D. degrees in 1982, 1986, and 1988, respectively, from the Massachusetts Institute of Technology, all in electrical engineering and computer science. He received the 1993 National Science Foundation Presidential Faculty Fellows Award and was also honored as Maryland's 1997 Outstanding Young Engineer. He is currently the William B. Kouwenhoven Professor of Electrical and Computer Engineering and holds joint appointments in the Departments of Radiology and Biomedical Engineering and a secondary appointment in Applied Mathematics and Statistics.

REFERENCES

[1] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis I: Segmentation and surface reconstruction," *Neuroimage*, vol. 9, no. 2, pp. 179–194, 1999.

[2] X. Han, D. L. Pham, D. Tosun, M. E. Rettmann, C. Xu, and J. L. Prince, "CRUISE: Cortical reconstruction using implicit surface evolution," *Neuroimage*, vol. 23, no. 3, pp. 997–1012, 2004.

[3] H. A. Drury, D. C. V. Essen, C. H. Anderson, W. C. Lee, T. A. Coogan, and J. W. Lewis, "Computerized mapping of the cerebral cortex: A multiresolution flattening method and a surface-based coordinate system," *J. Cogn. Neurosci.*, vol. 8, no. 1., pp. 1–28, 1996.

[4] T. Y. Kong and A. Rosenfeld, "Digital topology: Introduction and survey," *Comput. Vis., Graph., Image Process.*, vol. 48, no. 3, pp. 357–393, 1989.

[5] S. Li, *Markov Random Field Modeling in Image Analysis*. New York: Springer-Verlag, 2009.

[6] P.-L. Bazin and D. Pham, "Topology-preserving tissue classification of magnetic resonance brain images," *IEEE Trans. Med. Imaging*, vol. 26, no. 4, pp. 487–496, 2007.

[7] P.-L. Bazin and D. Pham, "Homeomorphic brain image segmentation with topological and statistical atlases," *Med. Image Anal.*, vol. 12, no. 5, pp. 616–625, 2008.

[8] V. Singh, H. Chertkow, J. Lerch, A. Evans, A. Dorr, and N. Kabani, "Spatial patterns of cortical thinning in mild cognitive impairment and Alzheimer's disease," *Brain*, vol. 129, no. 11, pp. 2885–2893, 2006.

[9] L. Abrams and D. Fishkind, "A genus bound for digital image boundaries," *SIAM J. Discrete Math.*, vol. 19, no. 3, pp. 807–813, 2005.

[10] L. Euler, "Solutio problematis ad geometriam situs pertinentis," in *Commentarii Academiae Scientiarum Petropolitanae*, vol. 8. 1741, pp. 128–140.

[11] K. Ueno, K. Shiga, T. Sunada, E. Tyler, and S. Morita, *A Mathematical Gift: The Interplay Between Topology, Functions, Geometry, and Algebra*. Providence, RI: American Mathematical Society, 2005.

[12] J. Foley, *Computer Graphics: Principles and Practice*. Reading, MA: Addison-Wesley Professional, 1995.

[13] U. Grenander and M. Miller, *Pattern Theory: From Representation to Inference*. New York: Oxford Univ. Press, 2007.

[14] P.-L. Bazin, L. Ellingsen, and D. Pham, "Digital homeomorphisms in deformable registration," in *Proc. Int. Conf. Information Processing in Medical Imaging 2007 (IPMI'07)*, vol. LNCS 4584, pp. 211–222.

[15] G. Bertrand, "Simple points, topological numbers and geodesic neighborhood in cubic grids," *Pattern Recognit. Lett.*, vol. 15, no. 10, pp. 1003–1011, Oct. 1994.

[16] L. Robert and G. Malandain, "Fast binary image processing using binary decision diagrams," *Comput. Vis. Image Underst.*, vol. 72, no. 1, pp. 1–9, 1998.

[17] J.-F. Mangin, V. Frouin, I. Bloch, J. Regis, and J. L. pez Krahe, "From 3-D magnetic resonance images to structural representations of the cortex topography using topology preserving deformations," *J. Math. Imaging Vis.*, vol. 5, no. 4, pp. 297–318, 1995.

[18] N. Passat, C. Ronse, J. Baruthio, J. Armspach, C. Maillot, and C. Jahn, "Region-growing segmentation of brain vessels: An atlas-based automatic approach," *J. Magn. Reson. Imaging*, vol. 21, no. 6, pp. 715–725, 2005.

[19] Y. Zeng, D. Samaras, W. Chen, and Q. Peng, "Topology cuts: A novel min-cut/max-flow algorithm for topology preserving segmentation in n-d images," *Comput. Vis. Image Underst.*, vol. 112, no. 1, pp. 81–90, 2008.

[20] C. Xu, D. Pham, and J. Prince, "Image segmentation using deformable models," in *Handbook of Medical Imaging*, vol. 2, J. M. Fitzpatrick and M. Sonka, Eds. Bellingham, WA: SPIE, 2000, ch. 3, pp. 129–174.

[21] D. MacDonald, N. Kabani, D. Avis, and A. C. Evans, "Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI," *Neuroimage*, vol. 12, no. 3, pp. 340–356, 2000.

[22] X. Han, C. Xu, and J. Prince, "A topology preserving level set method for geometric deformable models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 6, pp. 755–768, June 2003.

[23] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3-D surface construction algorithm," in *Proc. SIGGRAPH'87*, 1987, vol. 21, pp. 163–169.

[24] Z. Wood, H. Hoppe, M. Desbrun, and P. Schroder, "Removing excess topology from isosurfaces," *ACM Trans. Graph.*, vol. 23, no. 2, pp. 190–208, 2004.

[25] X. Han, C. Xu, U. Braga-Neto, and J. L. Prince, "Topology correction in brain cortex segmentation using a multiscale, graph-based algorithm," *IEEE Trans. Med. Imaging*, vol. 21, no. 2, pp. 109–121, 2002.

[26] D. Shattuck and R. Leahy, "Automated graph-based analysis and correction of cortical volume topology," *IEEE Trans. Med. Imaging*, vol. 20, no. 12, Nov. 2001.

[27] A. Szymczak and J. Vanderhyde, "Extraction of topologically simple isosurfaces from volume datasets," in *Proc. IEEE Visualization*, Seattle, Oct. 2003, pp. 67–74.

[28] F. Segonne, E. Grimson, and B. Fischl, "Topological correction of subcortical segmentation," in *Proc. 6th Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI'03)*, Montreal, Nov. 2003, pp. 695–702.

[29] N. Kiegeskorte and R. Goebel, "An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes," *Neuroimage*, vol. 14, no. 2, pp. 329–346, 2001.

[30] F. Segonne, J. Pacheco, and B. Fischl, "Geometrically accurate topology-correction of cortical surfaces using nonseparating loops," *IEEE Trans. Med. Imaging* (Special Issue on Computational Neuroanatomy), vol. 26, no. 4, pp. 518–529, 2007.

[31] B. Fischl, A. Liu, and A. Dale, "Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex," *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 70–80, 2001.

[32] P.-L. Bazin and D. Pham, "Topology correction of segmented medical images using a fast marching algorithm," *Comput. Methods Programs Biomed.*, vol. 88, no. 2, pp. 182–190, 2007.

[33] L. Abrams, D. Fishkind, and C. Priebe, "The generalized spherical homeomorphism theorem for digital images," *IEEE Trans. Med. Imaging*, vol. 23, no. 5, pp. 655–657, 2004.

[34] G. E. Christensen, S. C. Joshi, and M. I. Miller, "Volumetric transformation of brain anatomy," *IEEE Trans. Med. Imaging*, vol. 16, no. 6, pp. 864–877, Dec. 1997.

[35] S. Faisan, N. Passat, V. Noblet, R. Chabrier, and C. Meyer, "Topology preserving warping of binary images: Application to atlas-based skull segmentation," in *MICCAI '08: Proc. 11th Int. Conf. Medical Image Computing and Computer-Assisted Intervention—Part I*, 2008, pp. 211–218.

[36] X. Fan, P.-L. Bazin, and J. Prince, "A multi-compartment segmentation framework with homeomorphic level sets," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition 2008, CVPR'08*, pp. 1–6.

[37] N. Shiee, P.-L. Bazin, A. Ozturk, P. A. Calabresi, D. S. Reich, and D. L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," *Neuroimage*, vol. 49, no. 2, pp. 1524–1535, 2010.

[38] G. Sundaramoorthi and A. J. Yezzi, "Global regularizing flows with topology preservation for active contours and polygons," *IEEE Trans. Image Processing*, vol. 16, no. 3, pp. 803–812, 2007.

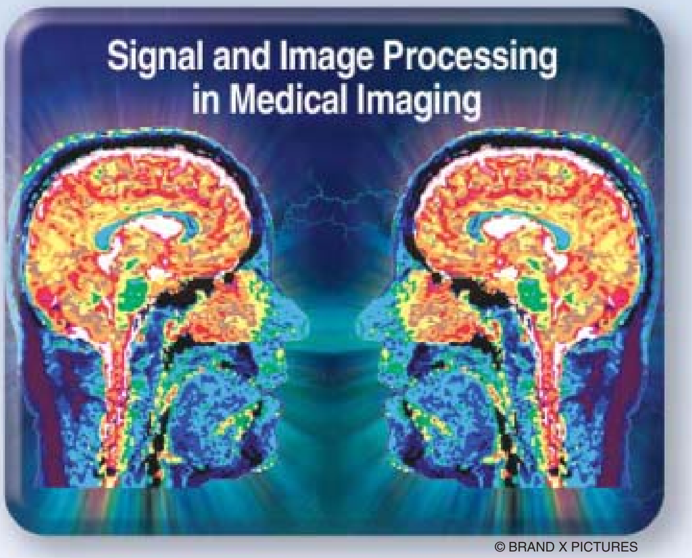
[39] F. Segonne, "Active contours under topology control—Genus preserving level sets," *Int. J. Comput. Vis.*, vol. 79, no. 2, pp. 107–117, 2008.

[40] Johns Hopkins University. Digital topology resources in human brain mapping [Online]. Available: http://medic.rad.jhu.edu/projects/digital_topology

Rolf Clackdoyle and Michel Defrise

Tomographic Reconstruction in the 21st Century

Region-of-interest reconstruction from incomplete data



© BRAND X PICTURES

Image reconstruction from projections is the field that lays the foundations for computed tomography (CT). For several decades, the established principles were applied not only to medical scanners in radiology and nuclear medicine but also to industrial scanning. When speaking of image reconstruction from projections, one is generally considering the problem of recovering some density function from measurements taken over straight lines, or the "line-integral model" for short. Image reconstruction can be performed by directly applying analytic formulas derived from the theory or by using general optimization methods adapted to handling large linear systems. The latter techniques are referred to as iterative to distinguish them from the analytic (or direct) methods. This article considers only the analytic methods. The two-dimensional (2-D) reconstruction problem (or classical tomography) refers to a density function in two dimensions with measurement lines lying in the plane, and the three-dimensional (3-D) problem considers 3-D density functions and lines with arbitrary orientations in space. The widely used term 3-D imaging is potentially confusing in this context, because there are some 3-D forms of image reconstruction that are mathematically equivalent to performing 2-D reconstruction on a set of parallel contiguous planes. To emphasize the distinction between 2-D and 3-D reconstruction, the terminology fully 3-D image reconstruction (or sometimes truly 3-D reconstruction) was introduced in the late 1980s when there seemed to be very little left to do in two dimensions but a rich unexplored 3-D theory to be developed.

INTRODUCTION

Classical tomography was, for the most part, developed in a research boom in the 1970s, with many publications on both iterative and analytic methods. During this boom, it was noticed that J. Radon had already solved the reconstruction problem in 1917 and that his solution was equivalent to the independently developed filtered backprojection (FBP) method. Fanbeam reconstruction methods were developed at the end of

Digital Object Identifier 10.1109/MSP.2010.936743

1053-5888/10/$26.00©2010IEEE

the 1970s, and interest then turned to the study of incomplete data problems, defined as problems where the line integral data are measured only for a proper subset of the set of lines crossing the object. These studies considered in particular truncated projections, limited-angle data, and exterior data problems, finally concluding in the late 1980s that incomplete data reconstruction always implied some kind of approximation in the image unless significant a priori information had somehow been incorporated.

By 1990, 2-D analytic image reconstruction was considered a mature field. It had produced the famous FBP algorithm for reconstruction from complete data, and incomplete data reconstruction would either require some good guesswork or would produce images with artifacts. There were few mathematical questions left to solve, and in the context of imaging in nuclear medicine at least, the line-integral model had already been largely superseded by more pertinent models that took into account physical effects in the scanners such as Compton scattering, self-attenuation, and noise due to photon-counting statistics. In the X-ray CT field, where line-integral models remained relevant, the field had started to move into the mathematically challenging (and fully 3-D) domain of cone-beam reconstruction that is still an active research area today. By the turn of the century, not only was 2-D reconstruction theory now completely understood, but after several decades of widespread application, the 2-D FBP algorithm was rapidly disappearing from the scene altogether, as medical and industrial scanners turned to fully 3-D reconstruction methods and/or iterative reconstruction schemes.

Suddenly, in 2002, to the astonishment of the research community, the first examples appeared of accurate partial reconstructions in two dimensions from incomplete data. These examples contradicted the understanding that incomplete data must inevitably generate artifacts throughout the image, and it then became important to distinguish between incomplete data (not all lines are measured) and insufficient data (not allowing accurate reconstruction in a given region of interest). The widespread belief that 2-D tomography was “all or nothing” had been shattered, and new mathematical problems arose as well as the potential for exciting new applications. Reduced measurement requirements immediately suggested important dose reductions for some kinds of scans. Also, oversize specimens [such as very large patients that exceed the scanner field of view (FOV)] could be at least partially imaged. Limited angular scans that admit accurate partial reconstructions could have significant implications in industrial applications.

In the new paradigm, only a subset of the full data was now needed to perform region-of-interest (ROI) reconstruction. This phenomenon had long since existed in some form in fully 3-D image reconstruction but was thought to be impossible in the 2-D case. In the context of 3-D reconstruction, this new 2-D capability introduced the possibility of some kinds of transverse data truncation in cone-beam scanning. The challenge now was to reconcile these new 2-D partial data reconstructions with the existing theory of the 1980s and to establish clearly under which data measurement conditions ROI reconstruction could produce truly quantitative and reliable images. The purpose of this article is to describe some of these recent advances in accurate 2-D ROI reconstruction from partial data and to resolve the apparent contradictions between these new methods and previous understanding.

[IMAGE RECONSTRUCTION CAN BE PERFORMED BY DIRECTLY APPLYING ANALYTIC FORMULAS DERIVED FROM THE THEORY OR BY USING GENERAL OPTIMIZATION METHODS ADAPTED TO HANDLING LARGE LINEAR SYSTEMS.]

20TH CENTURY STATE OF THE ART

We summarize here the state of the art of 2-D image reconstruction theory at the turn of the century. More details on the concepts and mathematical demonstrations can be found in various textbooks [1]–[4] or review articles (e.g., [5]–[6]). Iterative reconstruction methods are not covered in this article; we refer the reader to [7]–[9] for an overview.

FILTERED BACKPROJECTION

We first establish some notational conventions. The variables α and β will always represent unit vectors in the plane, whose directions are given by ϕ as follows:

$$\alpha = (\cos\phi, \sin\phi) \tag{1}$$
$$\beta = (-\sin\phi, \cos\phi). \tag{2}$$

The unknown density function will be denoted $f(x) = f(x_1, x_2)$, and the projection data will be denoted $p(\phi, s)$, which is the line integral of the density function along the line oriented at angle ϕ from the horizontal (x_1) axis and at a signed distance s from the origin. So

$$p(\phi, s) = \int_{-\infty}^{\infty} f(r\alpha + s\beta)dr \quad \text{for } \phi \in (0, \pi),\ s \in (-\infty, \infty); \tag{3}$$

see Figure 1. The one-dimensional function $p(\phi, \cdot)$ is called the parallel projection of the function f in the direction ϕ. It is rare that a scanner would collect line-integral data in this form of parallel projections, with uniform angular increments and uniform detector increments. However, it is always possible to represent the 2-D measurements in this format, and some algorithms start by interpolating the data into uniformly sampled parallel projections. When presented as a 2-D image, the data $p(\phi, s)$ is called the sinogram. Equation (3) expresses the 2-D Radon transform, which maps the density function f to the sinogram p. In this article, we assume arbitrarily fine sampling of the variables s and ϕ within their measured domains defined by the scanning geometry.

We note that the measurement is supposed to be over an infinite line as indicated by the integration limits of $-\infty$ to ∞. Implicit in (3) is the idea that the density function is zero outside some bounded region, so the integration really only takes place over this region. For a medical scanner, the patient port of the scanner is a suitable such region. We will use the term scanner port to indicate the region inside which the object lies. The object support is the region of nonzero densities and is always within the scanner port. In reality, the limits of integration should be implicitly viewed as being over the scanner port rather than over the entire plane, and the linear measurements s as taken over the corresponding finite range.

SUDDENLY, IN 2002, TO THE ASTONISHMENT OF THE RESEARCH COMMUNITY, THE FIRST EXAMPLES APPEARED OF ACCURATE PARTIAL RECONSTRUCTIONS IN TWO DIMENSIONS FROM INCOMPLETE DATA.

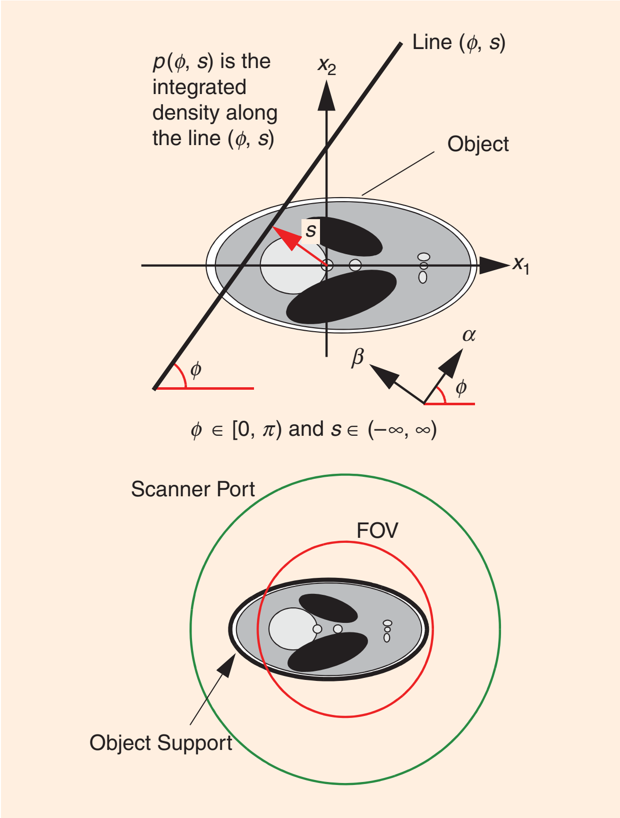
The FBP reconstruction formula is

$$f(x) = \int_0^\pi p_R(\phi, s)\big|_{s=x\cdot\beta}\, d\phi \qquad (4)$$

$$p_R(\phi, s) = \int_{-\infty}^{\infty} p(\phi, s')r(s-s')ds', \qquad (5)$$

where $r(s)$ is the ideal ramp-filter kernel whose Fourier transform is $R(\sigma) = |\sigma|$ so $p_R(\phi, \cdot)$ is called the ramp-filtered projection. The angular averaging to arrive at the reconstructed image in (4) is called the backprojection step. Note again in (5) that the integration limits are specified as $(-\infty, \infty)$ to avoid being specific about the finite extent of the projections. We use capital letters for Fourier transforms, so for example

$$R(\sigma) = \int_{-\infty}^{\infty} r(s)e^{-2\pi i\sigma s}ds, \quad r(s) = \int_{-\infty}^{\infty} R(\sigma)e^{+2\pi i\sigma s}d\sigma. \qquad (6)$$


Line (ϕ, s)
$p(\phi, s)$ is the integrated density along the line (ϕ, s)
x_2
Object
s
x_1
α
β
ϕ
ϕ
$\phi \in [0, \pi)$ and $s \in (-\infty, \infty)$
Scanner Port
FOV
Object Support

[FIG1] Some notation and terminology: (a) The variables ϕ and s are illustrated, with the unit vectors α and β. (b) The scanner field of view (FOV) is shown in red, and this object support is the black boundary. All the (mathematical) action takes place within the scanner port (green).

The formula of (4)–(6) has been known since the 1970s and follows from the central **section** theorem (also called the Fourier-slice theorem). The central-section theorem indicates the information contained in each projection; it relates the 2-D Fourier transform of the scanned object, to the one-dimensional Fourier transform of the projection

$$P(\phi, \sigma) = F(-\sigma \sin\phi, \sigma \cos\phi). \qquad (7)$$

Each transformed projection corresponds to a line of values passing through the origin (a central section) in the 2-D Fourier domain.

The FBP formula has also been shown to be equivalent to Radon's inversion formula. It is important to note that (in the appropriate mathematical context) the Radon transform is one-to-one [3]. This means that each (ideal, noise free) sinogram corresponds to a unique object. The FBP formula is an expression of the inverse transformation, taking sinograms back to density functions.

It will turn out that Fourier transforms have much less importance in partial-data problems. The Hilbert transform will play the central role, which we anticipate by rewriting the filtering part of the FBP formula. Note that $R(\sigma) = |\sigma| = (1/2\pi)(2\pi i\sigma)(-i\mathrm{sgn}\sigma)$, so we can express the ramp filter as a derivative composed with a Hilbert transform. Equation (5) can be replaced by

$$p_R(\phi, s) = \frac{1}{2\pi}\frac{\partial}{\partial s}p_H(\phi, s) \qquad (8)$$

$$p_H(\phi, s) = \int_{-\infty}^{\infty} p(\phi, s')h(s-s')ds' \quad \text{where} \quad h(s) = \frac{1}{\pi s} \qquad (9)$$

(recalling that in the Fourier domain, $H(\sigma) = -i\mathrm{sgn}\sigma$). The Hilbert transform performs a convolution with the function $1/\pi s$, and the integral of (9) is to be taken in the Cauchy principal value sense. Also, recall that in practice the integration takes place over a finite interval. We call p_H the Hilbert transform of the parallel projection.

Fanbeam projections also play a role in partial data problems. From a physical standpoint, fanbeam projections are more natural in the context of X-ray imaging, where the measurement rays all diverge from a point that corresponds to the location of the anode of the X-ray source. We refer to this point as the fanbeam vertex, v. The vertex follows a trajectory around the object, typically a circle outside the scanner port, but we will be more general here and parameterize the movement of the vertex as $v(\lambda)$ (which we shorten to v_λ), with a scalar variable $\lambda \in \Lambda$ where Λ is an interval. We use g to represent fanbeam data:

$$g(v_\lambda, \phi) = \int_0^\infty f(v_\lambda + l\alpha)dl \quad \text{for } \lambda \in \Lambda,\ \phi \in (0, 2\pi), \quad (10)$$

(where α is given by (1) as usual). The one-dimensional function $g(v_\lambda, \cdot)$ is called a fanbeam projection.

In both (3) and (10) we are representing line-integral measurements. We note that the link between fanbeam and parallel projections can be expressed as

$$p(\phi, s) = g(v_\lambda, \phi) + g(v_\lambda, \phi + \pi) \quad \text{where} \quad s = v_\lambda \cdot \beta \quad (11)$$

but that one of the terms in the sum on the right hand side is zero because the vertex v_λ is taken to be outside the convex hull of the object support (or even outside the scanner port). The formula $s = v_\lambda \cdot (-\sin\phi, \cos\phi)$ expresses the fact that the vertex v_λ lies on the line (ϕ, s).

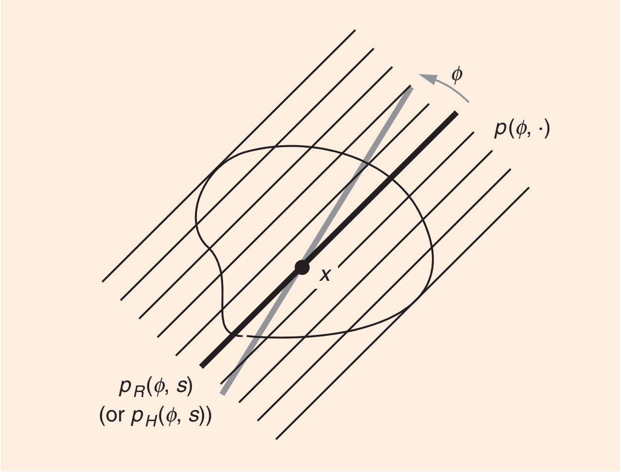
For a circular trajectory of the vertex, $v_\lambda = (-R_v\cos\lambda, -R_v\sin\lambda)$ with vertex radius R_v, the FBP scheme of (4) and (5) can be reformulated into a convenient fanbeam FBP formula (e.g., [4] or [6]). This formula has been the basis of reconstruction algorithms on virtually all CT scanners during the last two decades of the 20th century

$$f(x) = \frac{1}{2}\int_0^{2\pi} \frac{1}{\|x - v_\lambda\|^2}[g_1(v_\lambda, \phi)]\bigg|_{\phi = \arg(x - v_\lambda)} d\lambda \quad (12)$$

$$g_1(v_\lambda, \phi) = \int_{\lambda - \pi/2}^{\lambda + \pi/2} R_v \cos(\phi' - \lambda) g(v_\lambda, \phi') r(\sin(\phi - \phi'))\, d\phi'. \quad (13)$$

In (12), the integration limits of $(0, 2\pi)$ mean that the fanbeam vertex performs a full 360° scan. The factor of 1/2 in front of this integral compensates for the fact that each line intersecting the scanner FOV is measured twice during the scan. The shortest scan that allows each line to be measured at least once is the well-known *shortscan* equal to "180° plus the fan angle" where the "fan angle" refers to the aperture of the FOV as seen from the vertex. In the shortscan reconstruction formula, the factor of 1/2 is replaced by a weight function to suitably balance the subset of lines that are measured twice in a shortscan [10].

An important observation from (4) and (5) [or equivalently from (12) and (13)] should be noted. First, the left-hand side indicates the outcome of the reconstruction at a single arbitrary point x (arbitrary but inside the scanner port). Now a study of the operations on the right hand side shows that all values of the sinogram (or of the measured fanbeam projections) are used in the reconstruction formula because the ramp kernel $r(s)$ is known to be nonzero almost everywhere. Figure 2 provides a visual description of this fact (for the case of parallel projections). *At each point in the reconstructed image, all nonzero elements of the sinogram make a nonzero contribution to reconstruction.* This property of Radon's inversion formula strongly suggests that any missing data will affect the whole image, independent of the algorithm used for reconstruction. The effects of missing data and how to compensate in practice depend on which data are missing. Considerable effort has gone into understanding the nature of the various incomplete data situations.



[FIG2] Schematic of the FBP reconstruction algorithm. To reconstruct f at the point x, the contributions of $p_R(\phi, s)$ must be calculated (here $s = x \cdot \beta$ selects the line passing through x as shown). All the measured lines (ϕ, s') in the ϕ-projection make a contribution, weighted by the appropriate ramp kernel value $r(s - s')$. Reconstruction requires $p_R(\phi, s)$ for all ϕ so the argument is repeated for each angle as ϕ increases. Thus FBP requires nontruncated projections $p(s, \phi)$ for all angles ϕ. When using Hilbert transforms instead of ramp-filtering, the same argument applies to $p_H(\phi, s)$ (using contributions $h(s - s')$) except that a derivative is applied before backprojection.

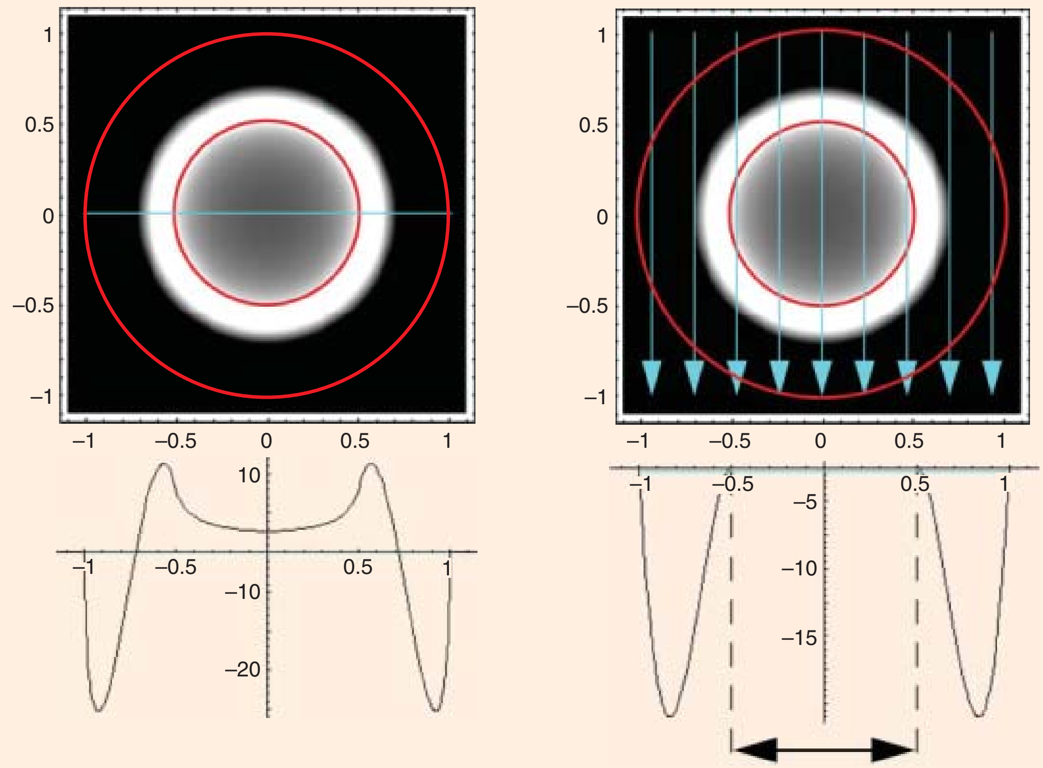
INCOMPLETE DATA

There is a large literature on incomplete data problems in classical tomography, mainly from the 1980s (e.g., [11]–[51]). There are two main types of incomplete data: truncated projections and limited-angle projections. A third category is exterior data, where the internal part of the projections are unavailable; this situation looks like a version of truncated projections, but actually falls into the category of limited-angle problems. We discuss these three situations in turn.

We say that a projection $p(\phi, \cdot)$ is complete (nontruncated) if for all s, $p(\phi, s)$ is either measured or known to be zero. Amongst the incomplete projections, truncated projections have the unmeasured lines at the extremities of the nonzero values. In terms of the FBP algorithm, the main difficulty with truncated projections is that the ramp filtering step to produce the filtered projections p_R from the measured projections p requires the entire projection. Using the equivalent derivative and Hilbert transform operations does not escape the difficulty as the Hilbert transform uses all elements of the projection. Most of the work in this area involved methods of extrapolating values into the truncated regions of the projections [13], [14], [32],[49], [50]. The case of full angular coverage but with all projections truncated on both sides is called the interior problem in the mathematics community. The objective is to reconstruct the region that is visible (nontruncated) in all projections, despite the contamination due to external parts of the object. It has been proven mathematically [52], [53], [3], [45] that the solution of the interior problem is not unique. Unambiguous reconstruction of the interior region-of-interest is impossible because multiple density functions can give the same measured

NONUNIQUENESS OF THE INTERIOR PROBLEM

Consider a true object f to be reconstructed, and some interior region A that is viewed in all the projections. The region A can be shown to be convex [56] and for simplicity we assume it to be circular of radius r_1. We move the origin of the system to the center of A, and note that since projections are truncated on both sides, there must be a second, larger circle that lies entirely inside the object; let r_2 be the radius of this second circle. Now it is possible to construct a function f_N such that i) $f_N(x) = 0$ for $\|x\| > r_2$, ii) $f_N(x) \neq 0$ (almost everywhere) for $\|x\| < r_1$, and iii) $p_N(\phi, s) = 0$ for all $\phi \in (0, \pi)$, $|s| < r_1$. Consider the measurements of the function f_N. The data are truncated at a radius r_1, and the measurements are zero inside this radius. However the function itself is nonzero. This means that both f and $f + f_N$ give the same measurements but are different almost everywhere inside the region of interest A. Obviously, the function f_N must contain both positive and negative densities, but even if it is known that the true function is nonnegative there will be object functions $f + kf_N$ (for a scalar constant k) which are nonnegative and give the same data as f. Note also that knowledge of the support of f does not improve the situation because the radius r_2 was chosen to keep the larger circle inside the object. Specific constructions of the function f_N can be found in [52], for example, and an example is shown in Figure S1 below. The construction of a suitable f_N requires all projections to be truncated on both sides [52, Th. 2].



Plot of density values of $f_N(x)$ along the x_1-axis.

Plot of projection values $p_N(\phi, s)$ for fixed ϕ.

[FIGS1] Illustration of a function f_N that is nonzero, but whose projection is zero in the interior (measured region). This example is circularly symmetric and is zero outside a circle of radius $r_2 = 1$. The interior region has a radius $r_1 = 0.5$. Any two objects that differ by f_N would produce identical interior data.

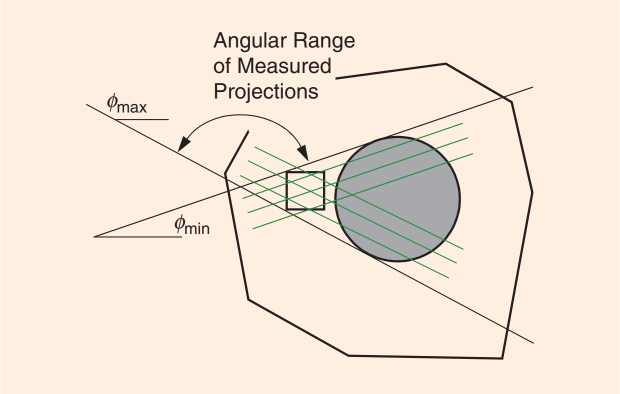
data. (Mathematically, the mapping from the object to the interior projections has a nontrivial nullspace; "Nonuniqueness of the Interior Problem" provides an element of the nullspace.)

On the other hand, for collections of nontruncated projections, uniqueness of partial data Radon transforms is relatively easy to achieve. A theorem in [52] (see also [3]) states that any infinite collection of nontruncated fanbeam or parallel projections has enough information to uniquely determine the object. However, uniqueness alone is not enough for effective image reconstruction, as we will see in the case of limited-angle data.

Although the term can be used more broadly, limited-angle data refers to the situation where a subset of parallel projections is not available. Specifically, a limited-angle data set is a collection of parallel projections $p(\phi, \cdot)$ for $\phi \in I$ where I is an interval or a union of intervals with total length (after performing modulo π) strictly less than π. From the central-section theorem (7), we immediately note that limited-angle data implies that a certain region of the 2-D Fourier transform F of the object f has not been directly measured. However, if the angular range I has nonzero length, the infinitely many projections along I ensure a unique solution (within a suitable class of object density functions, e.g., the L^2 functions on the scanner port) matching the limited-angle data. Uniqueness stems from the property that the Fourier transform F is an analytic function because f has bounded support. The various methods to uniquely recover the object density function f from limited-angle data are implicitly or explicitly based on analytic continuation, a process that is severely ill-posed or unstable. This instability is drastically more severe than the mild ill-posedness caused by the ramp filter in the FBP algorithm with complete data, and is more similar to the ill-posedness of problems such as the extrapolation of band-limited signals or the superresolution problems in imaging [54], [55]. Instability means that the inversion process is not continuous, so a minute error in the measurements can cause the reconstruction to jump to a solution far from the correct one (see "Stability and Instability in Image Reconstruction"). Reliable reconstruction is not possible in such a case, as attested by the limited success of the methods to treat limited-angle problems. Limited-angle data causes instability everywhere in the object because for nontruncated parallel projections, the pattern of measured lines is invariant throughout the object (because limited-angle systems are shift-invariant). In the sections "The Parallel-Fanbeam Hilbert Projection Equality" and "Differentiated Backprojection with Hilbert Filtering," we will see examples of problems with incomplete data for which instability or nonuniqueness only affects certain parts of the reconstructed image, and some partial region of interest can be recovered in a stable way from the incomplete data. Stability will be an important issue in the section "Further Advances Based on the DBP-H Approach."

The exterior problem in classical tomography refers to the situation where a central segment of all projections is unmeasured. The measured sinogram corresponds to the unmeasured part for the interior problem, so the interior and exterior problems are complementary in this sense. (Note that the objective is only to reconstruct the exterior region; there are no measurements whatsoever passing through the interior region. This is quite different from the situation of the interior problem.) For the exterior problem, Theorem 2 in [52] ([53, Th. 5.6]) applies, and unique reconstructions are possible. However, the exterior problem is unstable, because it is really a form of limited-angle tomography. As shown in Figure 3, consider any small region D to be reconstructed. There is a range of angles $[0, \phi_{\min}] \bigcup [\phi_{\max}, \pi]$ for which no measurement line passes through this region. We concentrate on reconstruction for just the region D. First we add some information: i) let us assume that the rest of the object density is known, and ii) let us add some hypothetical measurement lines to the small region such that complete projections $p(\phi, \cdot)$ of the region D are measured for all $\phi \in [\phi_{\min}, \phi_{\max}]$. Now since the rest of the object is known, its contribution to the measurements can be subtracted to leave measurements only of the isolated region D. We are now in the situation of (parallel projection) limited-angle tomography, which we know is unstable. So even by adding information we have an unstable problem in reconstructing this small region. As the region was chosen generally, we see that reconstruction of any part of the object is unstable for the exterior problem. A reconstruction from exterior data is shown in "Stability and Instability in Image Reconstruction."

Substantial literature exists on practical approaches to incomplete data problems, including both ad hoc methods of compensating for the missing information, and systematic theoretical approaches such as lambda tomography [57], [47]. These methods can produce useful reconstructions, especially when adapted to some specified imaging task. Here, however, we are concerned with reconstructions that are truly quantitative (and stable with respect to noise), in the same sense that FBP can produce arbitrarily accurate images under ideal implementation conditions (very fine discretization, high precision arithmetic)

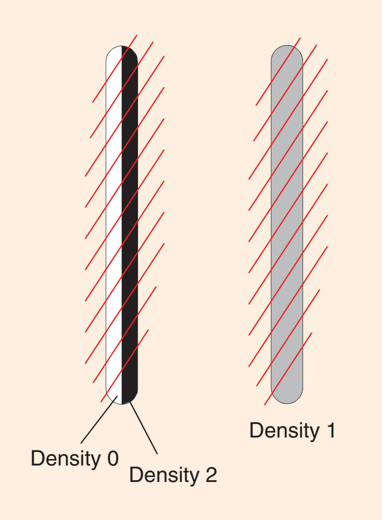


[FIG3] The exterior problem. All rays that do not pass through the shaded circle are measured. The objective is to reconstruct the outer ("exterior") region. The small square will only be measured for lines between angles $\phi_{\min}$ and $\phi_{\max}$. Consequently, adding some hypothetical measurements (lines shown in green, amongst others), and adding prior knowledge of the rest of the object then provides classical limited-angle data for the square, so it cannot be stably reconstructed, even though the reconstructed image is unique for the exterior problem.

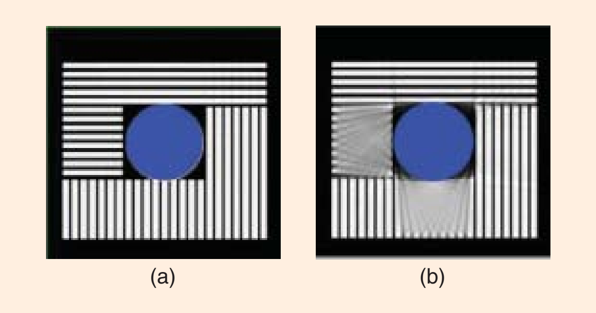
STABILITY AND INSTABILITY IN IMAGE RECONSTRUCTION

Image reconstruction from line integrals with complete or incomplete data is an ill-posed inverse problem. With such problems, proving that the solution is unique is not sufficient to guarantee that reliable reconstruction is possible in practice. Mathematically, uniqueness ensures the existence of an inverse operator, mapping the data to the solution, but this operator might be discontinuous implying that arbitrarily small perturbations of the measurements can cause arbitrarily large perturbations of the solution. For mildly ill-posed problems, stability is restored typically by restricting the class of admissible solutions using some a priori knowledge on the physical properties of the object being recovered. In most cases this prior knowledge reflects the expected smoothness of the solution. After proving uniqueness, it is thus essential in practice to investigate mathematically whether the reconstruction error, the difference between the solutions from noise-free and from noisy data, can be bounded in terms of the error on the measurements, assuming some reasonable constraint on the class of admissible solutions. Obtaining such error bounds is often straightforward when a closed form inversion formula is available, such as the FBP formula for reconstruction from complete data.

For image reconstruction problems, there exists a simple and practical operational method to identify situations where stable reconstruction is certainly impossible (identify that the problem is severely ill-posed). To establish instability of reconstruction at a point *P*, the method consists of searching for a line-segment centered at *P* that is not tangent to any measurement line. This line segment indicates an edge in the image that will be extremely difficult to recover—or rather, an edge that can't be reliably distinguished from other structures at that point in the image, as illustrated in Figure S2. This tangency requirement cannot be reversed to establish stability, but it does allow easy identification of incomplete data configurations where stable reconstruction is impossible, even though uniqueness may hold. The exterior problem is such a configuration, and is illustrated in Figure S3. This intuitive method is supported by a rigorous mathematical basis due to Finch [58]. A similar reasoning in the 3-D case [58] has been used to show that reconstruction from cone-beam projections is unstable at points where Tuy's condition [59] fails.



[FIGS2] Two objects, virtually indistinguishable from the measurements lines shown because none of the lines are tangent to the long edges.



[FIGS3] Reconstructions of the same phantom from (a) complete data and (b) from measurement lines not intersecting the blue circle (exterior data). Reconstructions were obtained by applying the ordered-subsets maximum likelihood expectation maximization (OSEM) algorithm (see e.g., [8]) to a discretized version of the problem (512 × 512 image, 512 × 512 sinogram, four subsets, 100 iterations). The exterior problem has a unique but unstable solution so accurate general reconstruction is impossible in practice. The bars that are poorly reconstructed have no measurement lines tangent to their long edges. Exterior data cannot stably recover all features of the object.

and ideal measurements (complete, finely sampled sinograms with arbitrarily low noise levels).

By the end of the 20th century, there were ample reasons to believe that stable reconstruction (even just ROI reconstruction) from incomplete data was impossible in classical tomography. First, a study of various incomplete data scenarios indicated that for a ROI to be stably reconstructed, that region must be measured from all angles. The interior problem satisfies this requirement, but the interior problem was shown to fail the uniqueness requirement. Therefore, neither limited-angle data nor truncated data can be tolerated for accurate stable reconstructions. The second and perhaps stronger evidence of complete data requirement for ROI reconstruction was the explicit mathematical form of the inverse transformation that maps sinograms back to objects [(4)–(5)]. As seen above, this inversion formula shows that each point in the object receives a contribution from every measured line through the object, even lines that pass far from the point being reconstructed (Figure 2). It simply didn't seem possible that ROI reconstruction could be performed with anything less than complete measurement data. In the sections "The Parallel-Fanbeam Hilbert Projection Equality" and "Differentiated Backprojection with

Hilbert Filtering," we describe two different mathematical approaches that successfully attacked this doctrine and established that stable ROI reconstructions from incomplete data could be achieved in certain situations.

THE PARALLEL-FANBEAM HILBERT PROJECTION EQUALITY

In this section, we present a mathematical formula with direct consequences for ROI reconstruction from incomplete (yet mathematically sufficient) data. We will see how it penetrates a slight gap in the analysis of partial data problems, and finesses the 'all or nothing' requirement of the FBP inversion formula.

THE HILBERT PROJECTION EQUALITY

First, the Hilbert transform of fanbeam data is defined

$$g_H(v_\lambda, \phi) = \int_0^{2\pi} g(v_\lambda, \phi')h(\sin(\phi - \phi'))\, d\phi' \qquad (14)$$

and we note that calculation of $g_H(v_\lambda, \phi)$ requires all values of the fanbeam projection $g(v_\lambda, \cdot)$, just as the calculation of $p_H(\phi, s)$ requires all values of the parallel projection $p(\phi, \cdot)$.

The parallel-fanbeam Hilbert projection equality [53] is

$$p_H(\phi, s) = g_H(v_\lambda, \phi) \quad \text{where} \quad s = v_\lambda \cdot \beta \qquad (15)$$

with p_H from (9). Recall from (11) that the condition $s = v_\lambda \cdot \beta$ means that the vertex point v_λ lies on the line (ϕ, s).

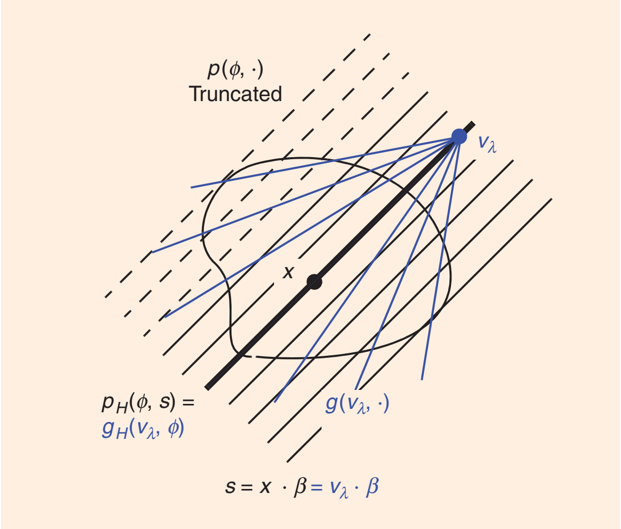
Equation (15) is of fundamental importance. It shows that there is some flexibility in obtaining $p_H(\phi, \cdot)$, in particular if any values of the $p(\phi, \cdot)$ projection are unavailable (truncated, for example). This is the key point, because it was previously assumed that $p_H(\phi, s)$ could not be obtained for any s if the projection $p(\phi, \cdot)$ was truncated. Instead, we see that $p_H(\phi, s)$ can be evaluated using a fanbeam projection provided (i) the fanbeam vertex v lies on the line (ϕ, s), and (ii) the fanbeam projection $g(v, \cdot)$ is not truncated; see Figure 4.

Demonstrations of (15) can be found in [61] and [60]. A full mathematical proof in a more general context appears in [53]. The simplified version of [60] is given in "A Demonstration of the Hilbert Projection Equality" on page 68.

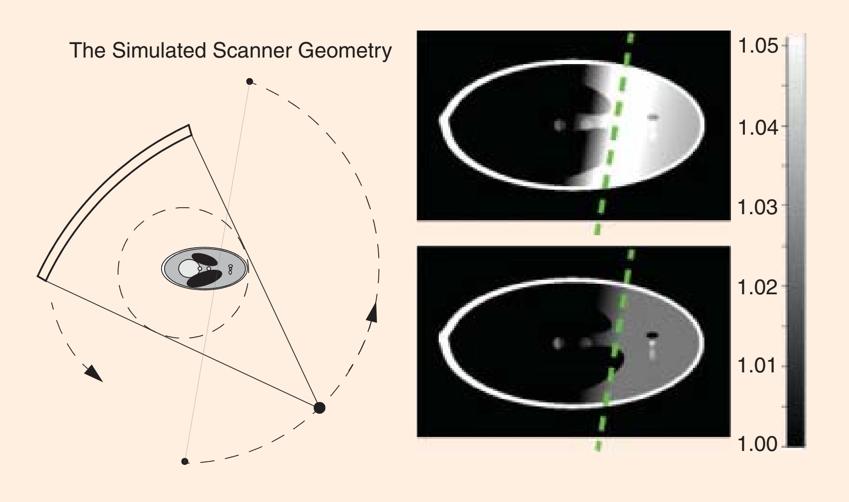
COMPLETE FANBEAM PROJECTIONS ON A REDUCED TRAJECTORY

A fanbeam projection $g(v, \cdot)$ is called complete (nontruncated) if $g(v, \phi)$ is measured or known to be zero for all ϕ. For the case of a 2-D scan consisting of complete fanbeam projections, the Hilbert projection equality provides a significantly improved data sufficiency condition, allowing partial reconstruction from a fanbeam trajectory on less than a shortscan, that is, from a fanbeam trajectory too short to measure all lines through the object (see Figure 5).

Fanbeam Data Condition: The point x can be reconstructed from complete fanbeam projections provided a fanbeam vertex can be found on each line passing through x [61]. (C1)



[FIG4] Implications of the Hilbert projection equality. A truncated parallel projection is shown (the dotted lines are unmeasured). According to parallel projection theory (see Figure 2), if any projection is truncated, then reconstruction at (any) point x cannot be performed because $p_H(\phi, s)$ cannot be obtained. However, the Hilbert projection equality shows that $p_H(\phi, s)$ might still be obtained via $g_H(v_\lambda, s)$ provided a complete (nontruncated) fanbeam projection $g(v_\lambda, \cdot)$ exists whose vertex lies on the line (ϕ, s). For data consisting entirely of complete fanbeam projections, the point x can be reconstructed provided a fanbeam vertex lies on each line passing through x.



[FIG5] Reduced (160°) fanbeam scan. Upper image: naive reconstruction (but nonetheless using standard fanbeam FBP with redundancy weighting $w(v_\lambda, \phi)$). Lower image: reconstruction using the Hilbert projection equality, in FBP format [(16) and (17)]. The reconstruction is accurate to the right of the dotted green line, where the fanbeam data condition (C1) is satisfied. (All simulations in this article use a modified Shepp-Logan phantom [62] with elongated outer ellipses, and adjusted intensities in the three small ellipses at right.)

A DEMONSTRATION OF THE HILBERT PROJECTION EQUALITY

We give a simplified demonstration (taken from [60]) of (15). We suppose that the line (ϕ, s) contains the vertex v (we drop the λ which plays no role in this discussion). We will translate the origin to v and rotate the axes by ϕ so that afterwards, $v = (0, 0)$, $\phi = 0$, and $s = v \cdot \beta = 0$. Now, consider

$$p_H(\phi, s) = \int_{-\infty}^{\infty} p(\phi, s')h(s - s')ds' = \int_{-\infty}^{\infty} p(0, s')h(0 - s')ds'$$

$$= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(r(1, 0) + s'(0, 1))dr \right\} h(-s')ds'$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{f(r, s')}{-\pi s'} drds'$$

and on the other hand

$$g_H(v, s) = \int_0^{2\pi} g(v, \phi')h(\sin(\phi - \phi'))d\phi'$$

$$= \int_0^{2\pi} g((0, 0), \phi')h(\sin(0 - \phi'))d\phi'$$

$$= \int_0^{2\pi} \left\{ \int_0^{\infty} f((0, 0) + l(\cos\phi', \sin\phi'))dl \right\} h(-\sin\phi')d\phi'$$

$$= \int_0^{2\pi}\int_0^{\infty} \frac{f(l\cos\phi', l\sin\phi')}{-\pi\sin\phi'} \frac{1}{l} l dl d\phi'$$

and the change from polar coordinates $(l\cos\phi', l\sin\phi')$ to rectangular (r, s') completes the demonstration.

Under this condition the reconstruction procedure simply uses (4), (8), and (15). Equation (4) requires calculation of $p_R(\phi, s)$ for each line (ϕ, s) passing through x. The filtered projection value $p_R(\phi, s)$ can be obtained from a derivative of $p_H(\phi, s)$; see (8), and (15) shows how $p_H(\phi, s)$ can be obtained from $g_H(v_\lambda, \phi)$ using a suitable fanbeam projection $g(v_\lambda, \cdot)$ that is assured to exist by the fanbeam data condition.

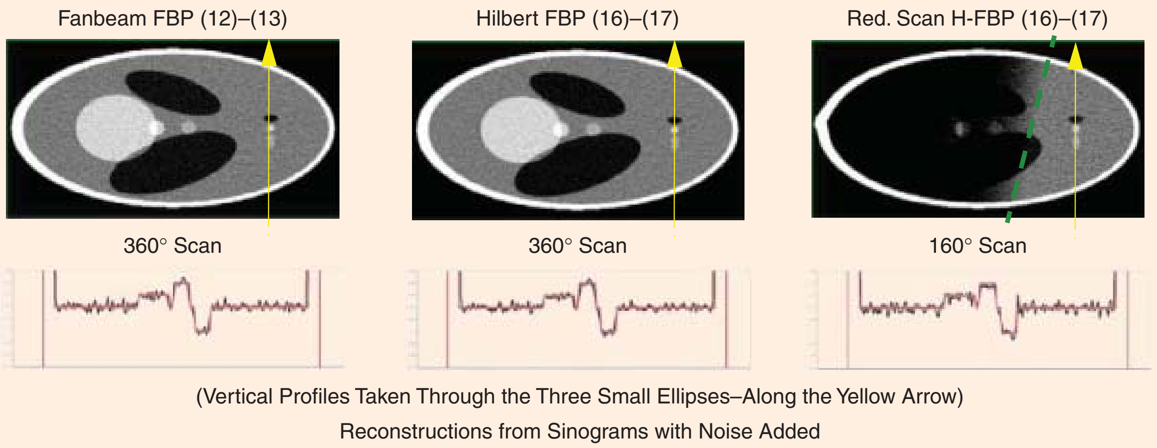
Aficionados of cone-beam reconstruction theory will recognize this fanbeam data condition as the exact 2-D analog of Tuy's ROI cone-beam data sufficiency condition [59], published in 1983. The Hilbert projection equality, (15), and higher-dimensional versions have been known since at least 1980 [53], but the implications for 2-D fanbeam ROI reconstruction were only published in 2002 [61].

Furthermore Noo et al. [61] extended this concept to produce new fanbeam formulas in the same FBP format as (12) and (13) but with the significant advantage that fanbeam data from reduced trajectories allow accurate reconstructions within ROIs that satisfy the data condition (C1). For the case of a (partial) circular trajectory with $\lambda \in \Lambda \subseteq [0, 2\pi]$, the ROI reconstruction formula becomes

$$f(x) = \int_{\Lambda} \frac{1}{\| x - v_\lambda \|} [w(v_\lambda, \phi)g_2(v_\lambda, \phi)]\Big|_{\phi = \arg(x - v_\lambda)} d\lambda \quad (16)$$

$$g_2(v_\lambda, \phi) = \frac{1}{2\pi}\int_{\lambda - \pi/2}^{\lambda + \pi/2} h(\sin(\phi - \phi'))\, \frac{\partial}{\partial\lambda} g(v_\lambda, \phi')\, d\phi'. \quad (17)$$

[The notation $(\partial/\partial\lambda)g(v_\lambda, \phi')$ should be interpreted as $(\partial/\partial\lambda)\widetilde{g}(\lambda, \phi')$, where $\widetilde{g}(\lambda, \phi') = g(v_\lambda, \phi')$.] The weight function w provides compensation for the situation where a point x receives contributions from two different fanbeam projections [10]. For the case of a 360° scan, the natural choice is the constant $w = 1/2$. For a general circular scan consisting of collections of finite angular segments, a suitable choice would be $w(v_\lambda, \phi) = c(\lambda)/[c(\lambda) + c(\pi + 2\phi - \lambda)]$, where $c(\lambda)$ is a smooth function that is zero outside the segments of Λ and nonzero within Λ. A reconstruction from a 160° scan is illustrated in Figure 5.


Fanbeam FBP (12)–(13)
Hilbert FBP (16)–(17)
Red. Scan H-FBP (16)–(17)
360° Scan
360° Scan
160° Scan
(Vertical Profiles Taken Through the Three Small Ellipses–Along the Yellow Arrow)
Reconstructions from Sinograms with Noise Added

[FIG6] Comparison of noise behavior of reconstruction algorithms. The Hilbert projection equality FBP magnifies noise to roughly the same degree as does conventional FBP, even for the reduced 160° scan (which uses less data).

Note that for a 360° scan (with $\Lambda = [0, 2\pi]$ and $w(v_\lambda, \phi) = 1/2$), (16) and (17) do not collapse to the standard fanbeam FBP (12) and (13). In (16), the backprojection weighting term is $\|x - v_\lambda\|^{-1}$ whereas in (12) it is $\|x - v_\lambda\|^{-2}$. The filtered projection g_2 is formed quite differently from g_1. The two methods are not mathematically equivalent. The right-hand side (RHS) of (16) can only be shown to match the RHS of (12) by applying the definition of $g(v_\lambda, \phi)$, (10). This point has been mentioned in [63] and will be discussed in more detail in the section "Multiple Inversion Formulas for the 2-D Radon Transform." Reconstructions from a simulated 360° scan are shown in Figure 6, and they demonstrate that the algorithm of (16) and (17) is no more sensitive to noisy data than the existing method of (12) and (13). Performing a comparative digital implementation with similar noise levels is a convenient way to experimentally evaluate the stability of algorithms. When the noise of the reconstruction is very similar to that obtained with a different program (such as FBP), it suggests that both methods have similar stability properties, at least in practice. Other example reconstructions can be found in [61].

We end this section with a different formula for reduced fanbeam trajectories [63], that does collapse to (12) and (13) for 360° scans (with $w(v_\lambda, \phi) = 1/2$)

$$f(x) = \int_\Lambda \frac{1}{\|x - v_\lambda\|^2} \times [g_1(v_\lambda, \phi) w(v_\lambda, \phi) + g_3(v_\lambda, \phi) w'(v_\lambda, \phi]\Big|_{\phi = \arg(x - v_\lambda)} d\lambda \qquad (18)$$

$$g_3(v_\lambda, \phi) = \frac{R_v \cos(\phi - \lambda)}{2\pi} \int_{\lambda - \pi/2}^{\lambda + \pi/2} g(v_\lambda, \phi')\, h(\sin(\phi - \phi'))\, d\phi', \qquad (19)$$

where $w'(v_\lambda, \phi)$ means $(\partial/\partial\lambda)\widetilde{w}(\lambda, \phi)$ with $\widetilde{w}(\lambda, \phi) = w(v_\lambda, \phi)$, and g_1 is given by (13).

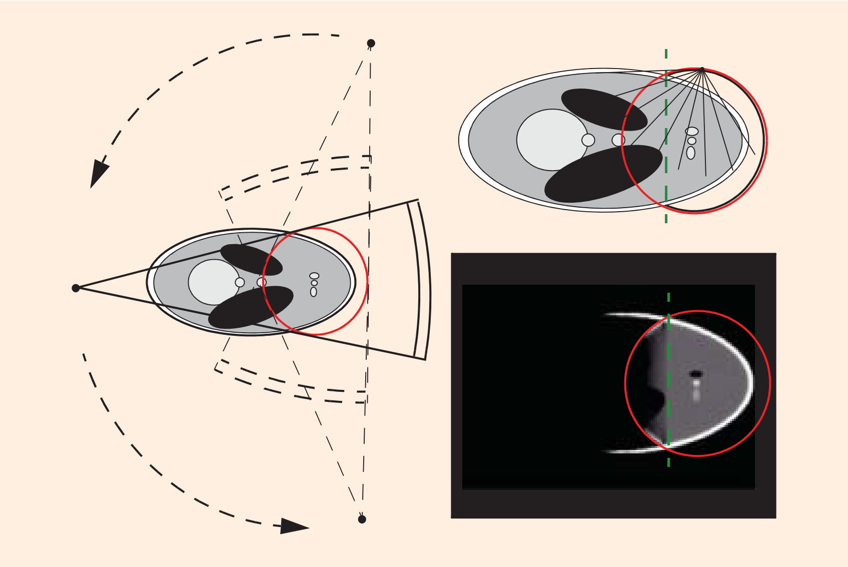
VIRTUAL FANBEAM PROJECTIONS

The Hilbert projection equality (15) can only be applied to nontruncated fanbeam projections. When the fanbeam vertices are far from the object, the projections become nearly parallel and the data condition for a point x to be reconstructed forces a nearly complete sinogram. It is the requirement of nontruncated projections that limits the application of this approach. This restriction can be considerably relaxed by considering virtual fanbeam projections [60]. The idea is not to look at how the projections were measured, but to consider rearrangements of the measured lines that form nontruncated fanbeam projections. We call these rearranged projections, virtual fanbeam (VFB) projections.

In the example of Figure 7, a fanbeam scan is taken with a circular vertex trajectory and a detector too small to cover the object. All the measured fanbeam projections are truncated, and some are even truncated on both sides. A substantial subset of the data can be rearranged into VFB projections by selecting (for example) VFB vertices inside the scanner FOV. These VFB projections are immediately seen to be nontruncated because all lines passing through the FOV have been measured. Many other valid (nontruncated) VFB projections outside the FOV are possible in this example by considering vertices to the right of the FOV. From the VFB trajectory shown, all points to the right of the green dashed line satisfy the data condition (C1) and can be reconstructed by (4), (8), and (15) as described in the section "Complete Fanbeam Projections on a Reduced Trajectory."

VFB vertices cannot be taken inside the object. To see why, note that the Hilbert projection equality requires half-line integrals from fanbeam vertices. For virtual vertices inside the object, the measurement line cannot be separated into two half lines. For virtual vertices outside the object, the required half line is extracted according to (11) with one of the terms on the right hand side being zero. The half-line integrals can always be obtained if the VFB vertices are outside the convex hull of the object support.

For general incomplete data problems, the VFB can be applied only if suitable VFB projections can be found that satisfy the data condition (C1) for points inside the desired ROI. Once suitable VFB projections have been found, the algorithm can proceed as described earlier, using (4), (8), and (15). More implementation details can be found in [61, Sec. 4.1]. The VFB vertices are not required to form a



[FIG7] Virtual fanbeam reconstruction. All the measured fanbeam projections are truncated (sometimes on both sides). A virtual trajectory can be taken just inside the scanner FOV, the red circle, ensuring nontruncated virtual fanbeam projections and valid reconstruction to the right of the green dotted line.

smooth trajectory, but if they do then fanbeam FBP reconstructions can be performed from the virtual projections following the description in [61, Sec. 4.2]. The main problem is, given an incomplete sinogram and knowledge of the object support (or in the worst case, knowledge of a convex region that contains the object, such as the scanner port), what is the largest ROI that can be reconstructed using the VFB approach? This reduces to identifying a maximal ROI satisfying the data condition for all possible sets of valid VFB projections.

We outline a procedure for finding this maximal ROI and the corresponding virtual vertices from a general incomplete data set. We describe the procedure in terms of parallel-projection sinograms because they provide a convenient format with which to represent general data measurements. We make the following assumptions: i) the (incomplete) sinogram is in parallel projection format p, ii) the set of unmeasured values is known, and iii) a convex region Ω that contains the object support is known. We remark that the smallest possible Ω should be used because this maximizes the flexibility in selecting the virtual vertices and therefore the size of the region where accurate reconstruction is possible. We also remark that virtual vertices are not needed for lines belonging to nontruncated (parallel) projections. The main steps of the algorithm are the following:

1) Identify T, the set of angles for which the projection $p(\phi,\ \cdot)$ is not complete.

2) Identify the region of complete angular coverage, region C, defined to be the set of points x in the scanner port for which all lines passing through x are either measured or do not intersect Ω. Define the region A as $\Omega\ \cap$ region C.

3) For all lines (ϕ, s) that cross region A and with $\phi \in T$, determine the set $V(\phi, s)$ defined as the intersection of the line (ϕ, s) with the region C\A. It is easily seen that $V(\phi, s) = \{v \notin \Omega : v \cdot \beta = s, g(v,\ \cdot)$ is nontruncated (completely measured)$\}$ is the set of valid VFB vertices on the line (ϕ, s).

4) The maximal ROI is given by $\{x \in \text{region}\, A : V(\phi, x \cdot \beta) \neq \emptyset \text{ for all } \phi \in T\}$, which is the subset of region A such that a valid VFB vertex exists for all lines passing through the point x.

The procedure is to search for valid VFB vertices on all lines where they are needed. They are not needed on any unmeasured lines, nor on lines of a complete (nontruncated) parallel projection. This tedious procedure to establish where the VFB method can be applied is a serious drawback to the method. The DBP approach of the section "Differentiated Backprojection with Hilbert Filtering" admits a simpler analysis, based on the geometry of regions A and C.

The VFB method was introduced in [60]. More examples can be found there and in [64] and [65].

MULTIPLE INVERSION FORMULAS FOR THE 2-D RADON TRANSFORM

In the sections "Complete Fanbeam Projections on a Reduced Trajectory" and "Virtual Fanbeam Projections," the Hilbert projection equality has been the key to reconstruction methods for some kinds of incomplete data problems. The very existence of these solutions demonstrates that multiple inversion formulas for the Radon transform must exist. If a certain ROI can be recovered from incomplete data, it can certainly be recovered from complete data. Since the FBP method for complete data uses all measured lines and the other method does not need certain lines, these two formulas must be nonequivalent. This important concept will be clarified in the section "Nonequivalent Reconstruction Formulas."

For reconstruction of the whole object using complete data, we have already seen two explicit inversion formulas (at the end of the section "Complete Fanbeam Projections on a Reduced Trajectory"). We show here using the VFB method that there are infinitely many inversion formulas, all fundamentally different yet all performing the unique 2-D inverse Radon transformation.

The idea is simple: the value $p_H(\phi, s)$ can be obtained from $g_H(v, \phi)$ for any VFB vertex v lying outside the object and on the line (ϕ, s). Assuming the object support lies within a circle of radius r, we can write

$$f(x) = \frac{1}{2\pi}\int_0^{\pi} \frac{\partial}{\partial s} q(\phi, s, t)\bigg|_{s = x\cdot\beta} d\phi \tag{20}$$

where $q(\phi, s, t) = g_H(v, \phi)$ for $v = s\beta + t\alpha$, and where t is only restricted in magnitude $|t| > \sqrt{r^2 - s^2}$ to ensure the virtual vertex is outside the object. After converting back to the original parallel projections and changing variables to match standard FBP format, an explicit form for $q(\phi, s, t)$ can be given (see [66])

$$q(\phi, s, t) = \frac{t^2 + s^2}{t^2}\int_{-r}^{r} \frac{p(\phi + \delta, s')h(ws - s')}{w} ds', \tag{21}$$

where the quantities δ and w depend on the variables $s,\ t,\ s'$ and are given by $\delta = \cos^{-1}(s'/\sqrt{s^2 + t^2}) - \text{sgn}(t)$ $\cos^{-1}(s/\sqrt{s^2 + t^2})$ and $w = \sqrt{s^2 + t^2 - s'^2}/t$. (The $\cos^{-1}$ function gives values in the range $[0, \pi)$ as usual.) As long as t is large enough, it can be chosen at leisure and it is allowed to vary with ϕ and s, so we can write $t_{\phi,s}$ to emphasize this dependence. Each choice of $t_{\phi,s}$ yields a different inversion formula that will correctly invert (3) and yet the formulas are not equivalent because each would react differently to noisy data. As $|t_{\phi,s}|$ increases, the virtual vertex on the line (ϕ, s) moves further from the object. In the limiting case of $|t_{\phi,s}| \rightarrow \infty$, formula (21) collapses to (9), and the reconstruction becomes standard FBP (for parallel projections). Note that t is not just a regularization parameter.

NONEQUIVALENT RECONSTRUCTION FORMULAS

The distinction between equivalent and nonequivalent reconstruction formulas concerns their behavior with respect to nonideal sinograms. Ideal sinograms are consistent with some object. Expressed mathematically, an ideal sinogram lies in the range of the 2-D Radon transform and satisfies certain range conditions, also called consistency conditions [3]. The principal aim of any reconstruction algorithm is to map these ideal sinograms back to the unique object consistent with it. The range conditions describe exactly what redundancies exist in an ideal sinogram, and they provide flexibility as to how the inverse mapping can be expressed, leading to nonequivalent inversion formulas. Nonequivalent inversion formulas give the same answer when presented with an ideal sinogram, but different images when presented with noisy sinograms [63], [67]. Equivalent inversion formulas, on the other hand, are simply mathematical rewritings of a particular inversion formula and would reconstruct the same images from noisy sinograms (except for implementation effects such as rounding errors, different regularizations, different discretizations, and sampling). The FBP formulas for parallel and fanbeam geometries presented in the section "20th Century State of the Art" [(4) and (5) versus (12) and (13)] are equivalent algorithms, whereas the fanbeam FBP formulas for 360° scans [(12) and (13) versus (16) and (17)] are nonequivalent. Two formulas are nonequivalent if it is impossible to transform one into the other by mathematical operations applying the range conditions, i.e., without using the original link, (3), between the object and the sinogram. For each choice of $t_{\phi,s}$, (20) and

DIFFERENTIATED BACKPROJECTION AND HILBERT TRANSFORMS

Parallel Projections

Using the definition (3), the derivative of the parallel projection is $(\partial/\partial s)p(\phi, s) = \int \beta\cdot\nabla f(s\beta + r\alpha)dr$. The backprojection of (26) is therefore

$$\bar{b}_{\phi_1,\phi_2}(x) = \frac{1}{\pi}\int_{\phi_1}^{\phi_2}\frac{\partial}{\partial s}p(\phi, s)\bigg|_{s=x\cdot\beta} d\phi$$
$$= \frac{1}{\pi}\int_{\phi_1}^{\phi_2}\int_{-\infty}^{\infty} \beta\cdot\nabla f((x\cdot\beta)\beta + r\alpha)dr\, d\phi.$$

Using $r' = r - x\cdot\alpha$, noting that $x = (x\cdot\beta)\beta + (x\cdot\alpha)\alpha$, and commuting the order of integration,

$$\bar{b}_{\phi_1,\phi_2}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{\phi_1}^{\phi_2} \beta\cdot\nabla f(x + r'\alpha)d\phi\, dr'.$$

Note that $d\alpha/d\phi \ldots \beta$, so $1/r' \ldots f(x + r'\alpha) = (\beta\cdot\alpha)$ and the integral over ϕ reduces to two boundary terms. Now recalling (22) and (9),

$$\bar{b}_{\phi_1,\phi_2}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{r'}(f(x + r'\alpha_2) - f(x + r'\alpha_1))dr'$$
$$= H_{-\alpha_2}f(x) - H_{-\alpha_1}f(x)$$

so $\bar{b}_{\phi_1,\,\phi_2}(x) = H_{\alpha_1}f(x) + H_{-\alpha_2}f(x)$, as illustrated in Figure S4. For the case $\phi_2 = \phi_1 + \pi$, we have $\alpha_2 = -\alpha_1$ so $\bar{b}_{\phi_1,\phi_1+\pi}(x) = 2H_{\alpha_1}f(x)$, which establishes (27).

Fanbeam Projections

By maintaining the vertex trajectory v_λ outside the convex hull of the object, we can extend the integration limits in definition (10) to $(-\infty, \infty)$, so we can write

$$g_D = \frac{\partial}{\partial\lambda}g(v_\lambda, \phi) = \int_{-\infty}^{\infty} v'_\lambda\cdot\nabla f(v_\lambda + l\alpha)dl,$$

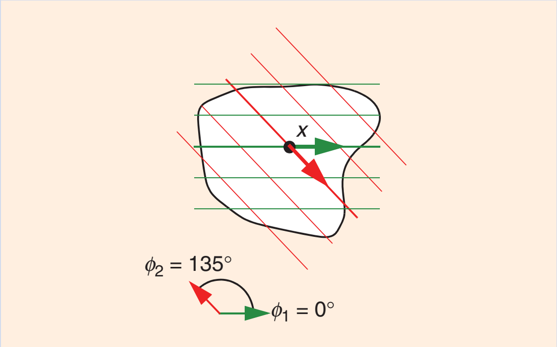
where $v'_\lambda = dv_\lambda/d\lambda$. Now the backprojection of g_D, given by (29), becomes

$$\hat{b}_{\lambda_1,\lambda_2}(x) = \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2}\frac{1}{\|x - v_\lambda\|}\int_{-\infty}^{\infty} v'_\lambda\cdot\nabla f\left(v_\lambda + l\frac{(x - v_\lambda)}{\|x - v_\lambda\|}\right)dl d\lambda,$$

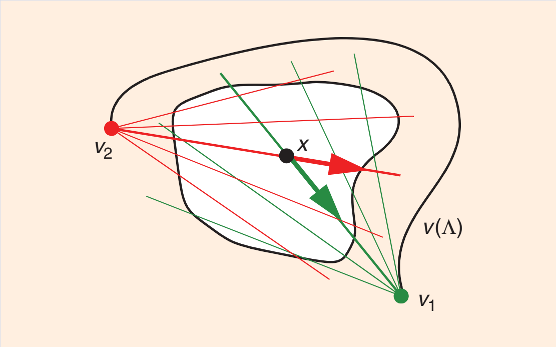
where we have used that $\alpha = y/\|y\|$ if $\phi = \arg y$. Now substituting $l = (1 - t)\|x - v_\lambda\|$, and then changing the order of integration after noting that $(d/d\lambda)f(x - t(x - v_\lambda)) = t\, v'_\lambda\cdot\nabla f(x - t(x - v_\lambda))$, the λ integral reduces to two boundary terms

$$\hat{b}_{\lambda_1,\lambda_2}(x) = \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2}\int_{-\infty}^{\infty} v'_\lambda\cdot\nabla f(x - t(x - v_\lambda))dt d\lambda$$
$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{t}(f(x - t(x - v_2)) - f(x - t(x - v_1)))dt$$
$$= H_{x-v_2}\, f(x) - H_{x-v_1}\, f(x)$$

so $\hat{b}_{\lambda_1,\lambda_2}(x) = H_{v_1-x}f(x) + H_{x-v_2}f(x)$ as illustrated in Figure S5. For $x \in (v_1, v_2)$ we have $v_1 - x = k_1(v_1 - v_2)$ and $v_2 - x = k_2(v_2 - v_1)$ for $k_1, k_2 > 0$, in which case $\hat{b}_{\lambda_1,\lambda_2}(x) = 2H_{v_1-v_2}f(x)$, which is (30), and from which (31) also follows easily.



[FIGS4] Illustration of $\bar{b}_{\phi_1,\,\phi_2}(x) = H_{\alpha_1}f(x) + H_{-\alpha_2}f(x)$.



[FIGS5] Illustration of $\hat{b}_{\lambda_1,\lambda_2}(x) = H_{v_1-x}f(x) + H_{x-v_2}f(x)$.

(21) generate a new nonequivalent inversion formula for the Radon transform, which illustrates the vast flexibility provided by the 2-D consistency conditions.

More importantly, this concept of nonequivalent inversion formulas explains why the argument at the end of the section “20th Century State of the Art” (that conventional FBP requires complete sinograms) does not, after all, preclude ROI reconstruction from incomplete data. Other, nonequivalent inverses might not draw on the whole sinogram to perform accurate reconstruction of some ROI. Several examples have been presented in this section, and the section “Differentiated Backprojection with Hilbert Filtering” will introduce a different approach, further increasing the range of incomplete sinograms that can be accurately handled.

DIFFERENTIATED BACKPROJECTION WITH HILBERT FILTERING

The Hilbert projection equality has opened the door to a number of results in ROI reconstruction from incomplete data, but they can only be applied for specific situations where enough complete fanbeam measurements or complete virtual fanbeam measurements are available to satisfy the data condition. In this section, we explore another class of incomplete data methods based on differentiated backprojection (DBP) with Hilbert post-filtering. In broad terms, these DBP-H methods operate in two steps: the first step is to perform a backprojection of the derivative of the projection data (called “DBP” to match the established “filtered backprojection” terminology), and the second step involves a post processing of this backprojected image involving the Hilbert transform. The method is easier to apply than the Hilbert projection equality methods that involve virtual fanbeam projections, and is more convenient for analyzing incomplete data problems. However, it too has restrictions on the kinds of incomplete data problems it can resolve.

The link between the Hilbert transform and the DBP is based on a general mathematical result of Gelfand and Graev [68]. The potential application to tomography was identified in 2002 by Finch [69], and made explicit by Noo et al. [70], Zhuang et al. [71], and Zou et al. [72] for parallel projections and for fanbeam projections. We only consider 2-D problems in this article, but the key ideas were developed simultaneously in three dimensions, providing elegant solutions to the cone-beam reconstruction problem in various configurations. See, for instance, [73]–[76] among many papers on this topic.

HILBERT IMAGES

The Hilbert transform operates on one-dimensional functions, but we can consider the Hilbert transform of an object function by applying one-dimensional transformations along parallel lines in a fixed direction. Thus we define an image $H_\alpha f$ obtained by performing Hilbert transforms along the direction α

$$H_\alpha f(x) = \int_{-\infty}^{\infty} f(x - t\alpha)h(t)dt \tag{22}$$

recalling that $h(t) = 1/\pi t$ and that the singularity at $t = 0$ is handled in the principal value sense. The object has finite support so the integration is performed over finite limits not explicitly specified. The Hilbert filtered image will nonetheless extend infinitely in the $+\alpha$ and $-\alpha$ directions. We note that $H_{-\alpha}f(x) = -H_\alpha f(x)$, and that we can extend the definition to arbitrary nonzero vectors v by

$$H_v f(x) = H_{v/\|v\|} f(x) = \int_{-\infty}^{\infty} f\left(x - t\frac{v}{\|v\|}\right)h(t)dt = \int_{-\infty}^{\infty} f(x - tv)h(t)dt \tag{23}$$

so H_v will apply a Hilbert filtering in the direction of the vector v; the magnitude of v is irrelevant.

Performing a Hilbert transform corresponds to multiplication by $-i\,\mathrm{sgn}\sigma$ in the Fourier domain, so applying a second Hilbert transform will, up to a minus sign, return the original function. Thus for any nonzero vector v,

$$H_v H_v f = -f. \tag{24}$$

We will see below that the result of the DBP operation will provide us with a certain Hilbert image that then needs to be inverted to complete the reconstruction. Unfortunately, (24) can only be applied if the $H_v f$ is known everywhere on its infinite extent. Aside from the difficulty of storing such a function, it will turn out that only part of the Hilbert image $H_v f$ can be obtained when considering partial data problems. Finding a suitable inverse to use instead of (24) will be the crux of the DBP-H method.

DIFFERENTIATED BACKPROJECTION

From now on, we extend the notation given by (1) and (2) in the obvious way: $\alpha' = (\cos\phi', \sin\phi')$, $\alpha_n = (\cos\phi_n, \sin\phi_n)$, and so on. We also define $\phi_0 = 0$, so $\alpha_0 = (1, 0)$ and $\beta_0 = (0, 1)$.

We consider parallel projections and fanbeam projections in turn. For the case of parallel projections, the derivative is taken with respect to the variable s

$$p_D(\phi, s) = \frac{\partial}{\partial s} p(\phi, s), \tag{25}$$

followed by a backprojection over the angular range (ϕ_1, ϕ_2) to yield

$$\bar{b}_{\phi_1,\phi_2}(x) = \frac{1}{\pi}\int_{\phi_1}^{\phi_2} p_D(\phi, s)\Big|_{s = x\cdot\beta} d\phi. \tag{26}$$

Straightforward mathematical manipulations (see “Differentiated Backprojection and Hilbert Transforms” on page 71) show that

$$\bar{b}_{\phi,\phi+\pi}(x) = 2H_\alpha f. \tag{27}$$

In particular, we note that $\bar{b}_{0,\pi} = 2H_{\alpha_0} f$, which means that backprojecting the derivative of the projections over the angular range $(0, \pi)$ results in the Hilbert transformed image in the x_1 direction (the horizontal, α_0 direction). Since $\bar{b}_{0,\pi}$

only involves a derivative and a backprojection, we see that $H_{\alpha_0}f(x)$ can be obtained if all lines passing through and near x are measured.

For fanbeam projections, we consider a smooth connected vertex trajectory $v(\lambda) = v_\lambda$ parameterized by λ, for $\lambda \in \Lambda = [\lambda_1, \lambda_2]$. The trajectory is located outside the convex hull of the object support. We define the derivative of the fanbeam projections by

$$g_D(v_\lambda, \phi) = \frac{\partial}{\partial \lambda} g(v_\lambda, \phi), \tag{28}$$

[where, as in (17), the RHS should be interpreted as $(\partial/\partial\lambda)\tilde{g}(\lambda, \phi)$ with $\tilde{g}(\lambda, \phi) = g(v_\lambda, \phi)$]. The backprojection of these differentiated fanbeam projections is given by

$$\hat{b}_{\lambda_1,\lambda_2}(x) = \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2} \frac{1}{\|x - v_\lambda\|} g_D(v_\lambda, \phi)\bigg|_{\phi = \arg(x - v_\lambda)} d\lambda. \tag{29}$$

To simplify the notations, we write v_k for v_{λ_k}, and we write $[v_1, v_2]$ to indicate the line segment joining v_1 to v_2. It can be shown that

$$\hat{b}_{\lambda_1,\, \lambda_2}(x) = 2H_{v_1 - v_2} f(x) \quad \text{if} \quad x \in [v_1, v_2]. \tag{30}$$

The direction of the Hilbert transform can be adjusted by breaking the trajectory into two segments at the point v_3, where $v_3 - x$ is the desired filtering direction

$$(\hat{b}_{\lambda_3,\lambda_2} + \hat{b}_{\lambda_3,\, \lambda_1})(x) = 2H_{v_3 - x} f(x) \;\; \text{if} \;\; x \in [v_1, v_2] \, \text{and} \, \lambda_3 \in (\lambda_1, \lambda_2); \tag{31}$$

see Figure 8.

Equation (31) describes a convenient method of obtaining $H_\alpha f$ directly from fanbeam projections rather than rearranging the data into parallel projections and applying (27). Even if the fanbeam trajectory consisted of several connected segments, $H_\alpha f(x)$ could be obtained using the fanbeam DBP of (28) and (29) for any x such that all lines through x intersect the trajectory. The backprojection would just need to be broken into the sum of suitable trajectory segments. We do not elaborate on this procedure here.

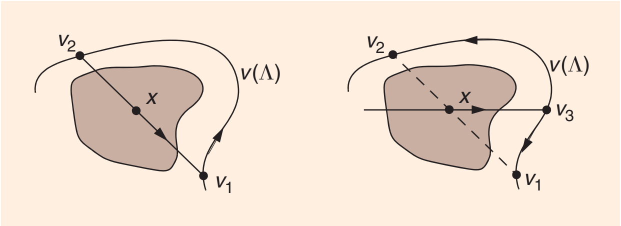
A different expression linking fanbeam data to the Hilbert transform can be considered, which involves $(\partial/\partial\phi)g(v_\lambda, \phi)$ instead of $(\partial/\partial\lambda)g(v_\lambda, \phi)$ in (29) [70]. This case will not be presented here.

INVERSION OF THE FINITE HILBERT TRANSFORM

If the Hilbert transform $Hq(s)$ of a function $q(s)$ is known for some interval $[L, R]$ that includes the support of the function q, then the function q can be recovered according to (see, for example, Section 4.3 (16) in [77])

$$q(t) = \frac{-1}{\sqrt{(t-L)(R-t)}}\left(\int_L^R \sqrt{(s-L)(R-s)}\frac{Hq(s)}{\pi(t-s)}ds - K\right) \tag{32}$$

for $s \in (L, R)$, and where the unknown constant K can be determined, for example, using the fact that $q(s_0) = 0$ for some known $s_0 \in (L, R)$. Alternatively, a calculation shows that $K = (1/\pi)\int q(s)ds$. Equation (32) is said to invert the finite Hilbert transform. In the context of image reconstruction, we sometimes refer to it as the truncated Hilbert transform in analogy with truncated projections, and also because the truncation of the Hilbert transform is related to truncation of the projections as we will see below.



[FIG8] DBP for fanbeam trajectories. If x lies on the line segment $[v_1, v_2]$, then $H_\alpha f(x)$ can be calculated by backprojecting g_D along i) the path (v_1, v_2) for α in the direction of $v_2 - v_1$, or ii) the paths (v_3, v_1) and (v_3, v_2) for α in the direction of $v_3 - x$.

Now in the image reconstruction context, we assume that the DBP (in either parallel or fanbeam format) produces an image $b_D(x)$ equal to $H_\alpha f(x)$. To simplify the explanations, we fix $\alpha = \alpha_0$ in this section, so $H_\alpha f$ refers to the Hilbert transform of the object density along horizontal lines. We will use the finite Hilbert inverse to reconstruct f along a fixed horizontal line say $x_2 = x_2^*$.

The object support along $x_2 = x_2^*$ is known to lie inside the interval $[f_L, f_R]$ and we assume that $b_D(x_1, x_2^*)$ is known for $x_1 \in [h_L, h_R]$ with $h_L < f_L < f_R < h_R$. In this case, we can apply formula (32) as follows: for all $x_1 \in (f_L, f_R)$,

$$f(x_1, x_2^*) = \frac{-1}{\pi\sqrt{(x_1 - h_L)(h_R - x_1)}} \int_{h_L}^{h_R} \sqrt{(s - h_L)(h_R - s)}\, b_D(s, x_2^*) \times \left(\frac{1}{(x_1 - s)} - \frac{1}{(\bar{x}_1 - s)}\right) ds, \tag{33}$$

where we have used $f(\bar{x}_1, x_2^*) = 0$ for some choice of $\bar{x}_1 \in (h_L, f_L) \cup (f_R, h_R)$.

It is important to note that the integration limits (h_L, h_R) in (33) encompass an interval for which the Hilbert transform is known and that contains the support of the unknown function: $(f_L, f_R) \subset (h_L, h_R)$. This requirement plays a role in the data condition for DBP-H reconstruction.

The procedure can be repeated for all rows of the DBP image (different values of x_2^*) to obtain a complete reconstructed image.

DBP-H RECONSTRUCTION

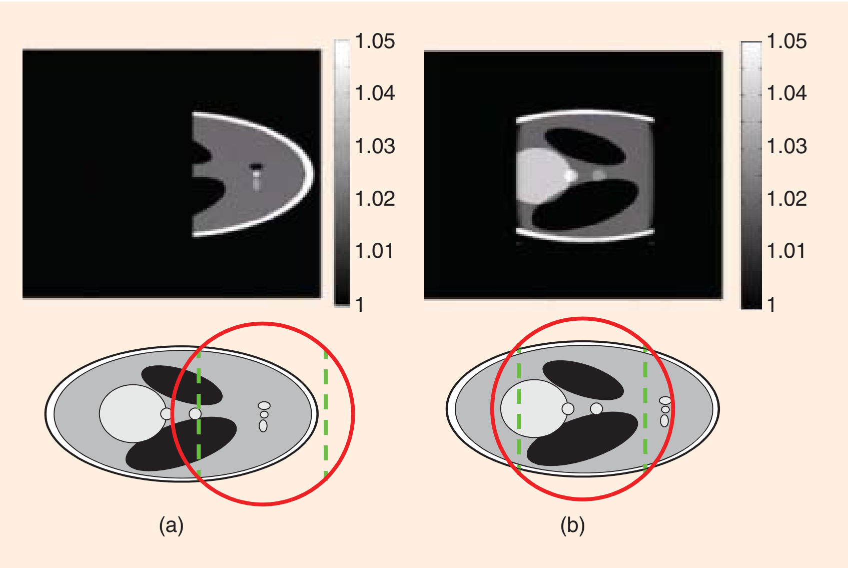
For complete data, whether in parallel or fanbeam format, the procedure for DBP-H reconstruction is to first obtain a DBP image over a larger region than the support of the object. From this image ($b_D(x)$), the finite Hilbert inverse, (33), can be applied for all horizontal lines to obtain the reconstructed image. Note that vertical lines could be used instead, or lines at any angle α. It can be shown that for any finite interval (h_L, h_R), this reconstruction method is not equivalent to standard FBP [(4) and (5) or (12) and (13)]. (Equivalence to FBP occurs in the limit as $(h_L, h_R) \to (-\infty, \infty)$.)

Figure 9 shows DBP-H reconstructions from incomplete data. The red circle indicates the FOV of the scanner; the set of measured lines are precisely those that pass through the red circle. So each point in this region of complete angular coverage has the property that all lines passing through the point are measured. In general, for such points x, the DBP $b_D(x)$ can be calculated for any choice of Hilbert filtering direction α (using (27) or (30) as appropriate). We see that in these cases (Figure 9), the appropriate filtering direction is vertical, so $\alpha = (0, 1)$. Thus the ROI reconstruction image was obtained by first forming a DBP image inside the red circle, with $\alpha = (0, 1)$. Then for each vertical line within the green boundaries, the inverse finite Hilbert transform was applied, according to (33) (but adjusted for vertical rather than horizontal lines). For this example, horizontal filtering would not have been possible, because along horizontal lines the support of f is not contained in the DBP region (i.e., the required condition $(f_L, f_R) \subset (h_L, h_R)$ is not satisfied).

For general incomplete sinograms, the procedure for ROI reconstruction using the DBP-H method is given below. The support of the object is assumed known, and it is known which line integrals are missing (unmeasured) from the sinogram. We first define the following three regions:

$$\text{Region } A = \{x \in \text{support } f : \text{all lines through } x \text{ are measured}\} \tag{34}$$

$$\text{Region } B = \{x \in \text{support } f : x \notin \text{region } A\} \tag{35}$$



[FIG9] Examples of ROI reconstructions using the DBP-H method. The ratio of the ellipse axes to the FOV radius is 1 : 2 : 4/3. (a) The object is positioned to the side of the FOV and (b) centered on the FOV.

$$\text{Region } C = \{x \in \text{scanner port} : \text{all lines through } x \text{ are measured or miss the object}\}. \tag{36}$$

The definitions of regions A and C are consistent with the earlier use in this article, and the definitions of regions A and B have previously appeared in the literature. Note also that region A is always a subset of region C, and that regions B and C do not intersect. The idea is that region C corresponds to the scanner FOV, which is the region for which an accurate DBP image can be formed. (However, strictly following the definition, region C depends on the object support also. For example, if the object is completely inside the FOV, then region C is the whole scanner port.)

The procedure for DBP-H reconstruction is the following:

1) From knowledge of the object support and the set of measured rays, determine the three regions A, B, C.

2) Identify reconstruction line segments. These segments intersect region A, have endpoints in region $C \backslash A$ (endpoints outside the object but still in region C), but do not intersect region B. A ROI reconstruction might be made up of several collections of parallel line segments.

3) For each reconstruction line segment, DBP [(27) or (30)] is performed to obtain $H_\alpha f(x)$ for α in the direction of the line, and all x on the line segment. Then the finite Hilbert inversion [(33)] can be performed because the line segment extends beyond region A to provide the small space between the endpoints of (f_L, f_R) and (h_L, h_R). The inverse finite Hilbert transform provides the reconstruction for all points along the line segment.

From this procedure follows the data condition for ROI reconstruction using the DBP-H method.

DPB-H Condition: The point x can be reconstructed if it lies on a line segment extending outside the object on both sides, and all lines crossing the line segment are measured. (C2)

The method reconstructs line segments using the finite Hilbert inverse, so it would be more natural to ask if a particular line segment can be reconstructed, rather than individual points. The reconstructed ROIs consist of unions of line segments that traverse the object without crossing region B.

It can be observed that the two ROI examples of Figures 5 and 7 can be reconstructed using the DBP-H approach. However, there are incomplete data examples for which data condition (C2) fails, and yet condition (C1) applies (see [70] and Figure 13). Furthermore, the example of Figure 9(b) cannot be handled using the VFB approach: not all points inside the reconstructed region satisfy data condition (C1); see [60] for an explanation.

FURTHER ADVANCES BASED ON THE DBP-H APPROACH

The sections "The Parallel-Fanbeam Hilbert Projection Equality" and "Differentiated Backprojection with Hilbert Filtering" presented two methods of analytic image reconstruction that apply to some kinds of incomplete data problems. Each method provides information about which ROIs can be reconstructed under particular incomplete data situations, and both methods can be written in the form of a single mathematical inversion formula. In this section, we describe recent theoretical advances in ROI reconstruction that do not provide explicit inversion formulas. These advances describe mathematical results that establish uniqueness and (more importantly) stability for certain ROI reconstruction problems. These mathematical results involve the inversion of some form of truncated Hilbert transform, which via the DBP formulation, implies a related inversion result for image reconstruction from truncated data. Although these inversion results do not provide direct analytic reconstruction algorithms, they are important for understanding the nature of 2-D ROI reconstruction, and do have implications for iterative reconstruction algorithms.

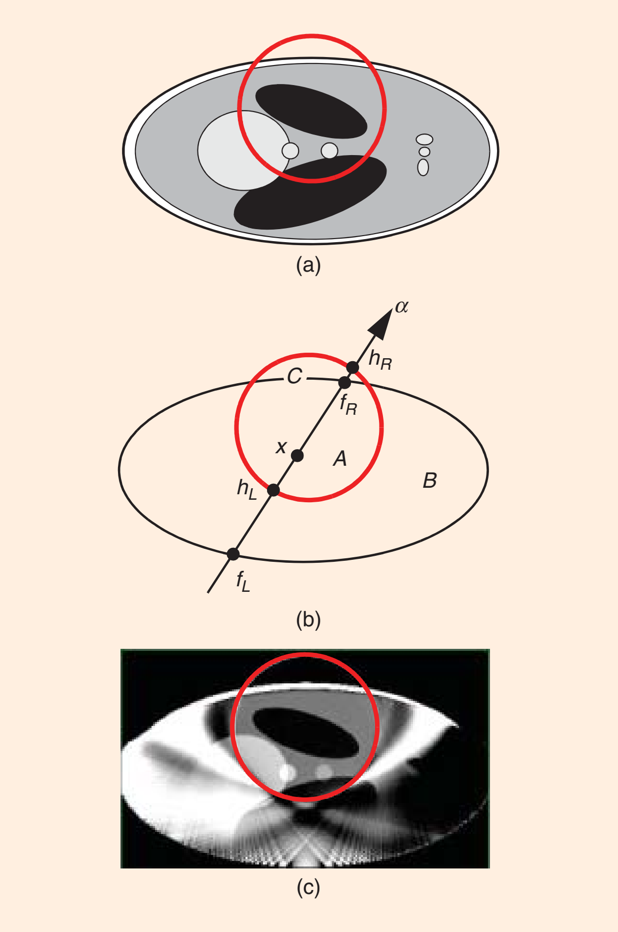
THERE ARE TWO MAIN TYPES OF INCOMPLETE DATA: TRUNCATED PROJECTIONS AND LIMITED-ANGLE PROJECTIONS.

INVERSION OF THE ONE-SIDED FINITE HILBERT TRANSFORM

An important advance in 2-D ROI reconstruction came from the work of Defrise et al. [78]. The motivation was the then unresolved case of a scanner FOV (region C) not crossing two boundaries of the object; see Figure 10. The DBP method discussed in the section "Differentiated Backprojection with Hilbert Filtering" could not be applied because only one end of a line segment contained in region C could lie outside the object, and therefore the conditions for the inverse finite Hilbert transform, formula (32), could not be satisfied. This configuration also defeats the VFB method because every point in the object has many lines that do not contain valid VFB vertices.

We begin by analyzing feasible regions of the object for potential ROI reconstruction. Note that the data conditions (C1) and (C2) (for VFB and DBP-H reconstruction, respectively) can only be satisfied for points inside region A. This fact follows directly from the way the methods are applied, but is also true for general incomplete data problems: the only possible points for unique stable reconstruction must lie inside region A. The reasoning is as follows. By definition, any point x inside region B lacks full angular coverage; there will be lines passing through x that are unmeasured. Then, following the argument used in the section "Incomplete Data" for the exterior problem, a small limited-angle reconstruction problem exists for a neighborhood of that point, and stable reconstruction at x is impossible. Thus stability cannot hold inside region B, and the only candidates for practical quantitative ROI reconstruction lie inside region A.

Uniqueness is relatively easy to establish. According to a theorem of K.T. Smith mentioned in [52, Sec. 2.2], any infinite collection of nontruncated fanbeam projections will ensure uniqueness. Such a collection can be found by assigning virtual fanbeam vertices



[FIG10] Example of an incomplete data problem that cannot be resolved using the VFB method (see the section "The Parallel-Fanbeam Hilbert Projection Equality") or the DBP-H method (see the section "Differentiated Backprojection with Hilbert Filtering"). (a) For a scanner FOV (shown in red) that is smaller than the width of the object, the only hope for reconstructing the two small central circles of the phantom would be to avoid the situation of the interior problem and to use a configuration such as illustrated. (b) The regions A, B, and C (34)–(36) are indicated. There are (almost) no line segments that traverse region A with both endpoints in region $C \setminus A$, so the DBP-H cannot be applied. For the VFB method, all valid virtual vertices must lie in region $C \setminus A$, but virtual vertices cannot be found on horizontal lines through region A. It can be shown that reconstruction is possible at a point x in region A by considering a special one-sided finite Hilbert transform along a line passing through x and the region $C \setminus A$. (c) Iterative reconstruction for this incomplete data problem, verifying the theoretical result of accurate reconstruction inside all of region A. (Reconstruction using OSEM: 512 × 512 image, 1,024 × 1,024 sinogram, 16 subsets, 20 iterations.)

along a trajectory in region $C \backslash A$. Thus we identify that the main issue for theoretically feasible ROI reconstruction is to establish stability.

The main result of [78] was the demonstration that the following finite Hilbert transform problem has a unique and stable solution. We recall that for a line passing through an object density function, the Hilbert transform $H_\alpha f(x)$ can be obtained from the DBP image $b_D(x)$, which can be computed for any point x in region A, and for any direction α. Fix some line passing through region A of the object and through region C outside the object. Taking the object support to be convex as usual, the density function along the line is known to be zero outside an interval $[f_L, f_R]$. The values of $H_\alpha f$ are known for a single interval $[h_L, h_R]$ corresponding to the intersection of the line with region A. Suppose now that $f_L < h_L < f_R < h_R$ as illustrated in Figure 10. We note that the interval $[h_L, h_R]$ no longer contains the object support $[f_L, f_R]$ as in the section "Differentiated Backprojection with Hilbert Filtering." It has been proved that $f(x)$ can be uniquely and stably recovered for $x \in (h_L, f_R)$ from the Hilbert transform values $H_\alpha f(x)$ for $x \in [h_L, h_R]$ [78]. The stability was established by showing that finite errors in the measurement values $H_\alpha f(x)$ do not produce infinite errors in the reconstructed values $f(x)$; see "Stability Estimates for Two Finite Hilbert Transforms."

The immediate consequence of this result is that unique, stable reconstruction is possible for all region A in Figure 10. An inversion formula is not currently known for this case, but the information is important for iterative algorithms that blindly search for a solution to the large discretized version of the linear system. Assuming that the discretized system approximates the continuous case, this ROI reconstruction result suggests which values of the iterative solution are reliable.

STABILITY ESTIMATES FOR TWO FINITE HILBERT TRANSFORMS

The One-Sided Finite Hilbert Transform

Let $f:\mathbb{R}\to\mathbb{R}$ be a smooth function that vanishes outside an interval $[f_L, f_R]$, and let the Hilbert transform $g = Hf$ be known for $x \in [h_L, h_R]$ where $f_L < h_L < f_R < h_R$. Furthermore, assume that the constant $K = (1/\pi)\int f(x)dx$ is known. Using analyticity arguments, it can be shown that f can be uniquely determined on the segment $(h_L, f_R]$. This is the basis for using the DBP with the configuration of Figure 10.

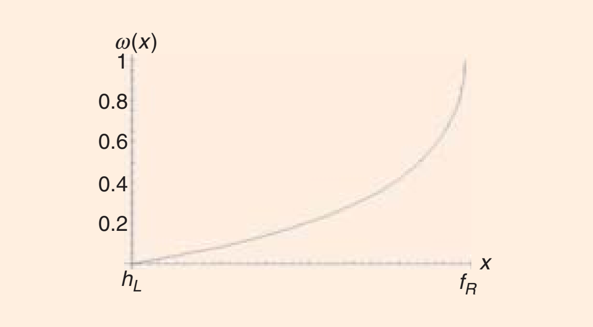
The practical significance of this result is limited if there is no control on the stability. Specifically, if the ideal "noise-free" Hilbert transform $g = Hf$ is replaced by the measured data $g^{\epsilon}(x) = g(x) + n(x)$ where $n(x)$ is the measured noise, one needs an upper bound on the reconstruction error $f^{\epsilon}(x) - f(x)$ where f^{ϵ} is the solution obtained from noisy data. Since no closed form inversion formula is known for this problem, stability cannot be analyzed directly. A stability estimate can nevertheless be obtained using prior information on the noise level and on the values of f or Hf in the unmeasured interval $[f_L, h_L]$. We set $-f_L = f_R = 1$ below to simplify the notations. In [78], it is assumed that a positive number M is known such that

$$\frac{1}{\pi}|g(x)|\sqrt{1-x^2} \le \frac{M}{2} \quad x \in [-1, h_L]$$

and that an upper bound ϵ is available for the measurement noise (see [78] for the precise definition of ϵ). With these assumptions, the measurement error is bounded by

$$\sqrt{1-x^2}|f^{\epsilon}(x) - f(x)| \le \epsilon + M\log\left(\frac{x+1}{x-h_L}\right) \times \left(\frac{2\epsilon}{M\log(2/(1-h_L))}\right)^{\omega(x)} \quad x \in (h_L, 1].$$

This equation specifies a minimum rate $\omega(x)$ at which the error tends to zero as a power of the noise ϵ. This power law dependence is important because the use of analyticity arguments to prove uniqueness might have suggested a worse convergence such as $1/\log\epsilon$. The rate of convergence is illustrated in Figure S6, [see [78] for an analytic expression for $\omega(x)$] which shows that $\omega(x)$ decreases as x moves away from the segment $[f_R, h_R]$ where f is known to be zero. The convergence is linear at the edge of that segment ($\omega(f_R) = 1$) but degrades as one moves away from the region where f is known a priori.



[FIGS6] The minimum error convergence rate.

The Interior Hilbert Transform, with Prior Knowledge

A similar stability estimate exists (see [81]) for the problem in the section "Inversion of the Interior Hilbert Transform," where f is known in a subset $[k_L, k_R]$ of the interior segment$[h_L, h_R] \subset [f_L, f_R]$ in which the Hilbert transform Hf can be recovered using the DBP. Assuming as above an upper bound M on $|Hf(x)|\sqrt{1-x^2}$ on the unmeasured segments$[f_L, h_L]$ and $[h_R, f_R]$ and an upper bound ϵ on the measurement noise, an upper bound on the reconstruction error in the segments $[h_L, k_L]$ and $[k_R, h_R]$ can be obtained, with a power law in $\epsilon^{\omega(x)}$, where $\omega(x)$ tends to zero as $x \to h_L$ or h_R, and $\omega(x)$ tends to one when x tends to the known segment, i.e., $x \to k_L$ or k_R.

However, this one-sided finite Hilbert transform stability result has consequences in a more general setting. We consider an arbitrary incomplete data problem for an object with known convex support. It has been pointed out [56] that every connected component of region A must be convex. The one-sided Hilbert transform result combined with the convexity of the object immediately shows that each convex component of region A can be uniquely and stably reconstructed provided it forms a proper subset of a connected component of region C (informally, provided the component of region A touches the boundary of the object, and that there is some of region C on the other side of the boundary). For the other connected components of region A that are internal to the object, it is not known to what extent reconstruction might be possible. Figure 11 illustrates this point with a highly artificial example.

Interest now turns to studying ROI reconstruction for (components of) region A internal to the object. In particular, for the interior problem one can ask how much information needs to be added to restore uniqueness and ensure stability. The next section provides one answer to this question.

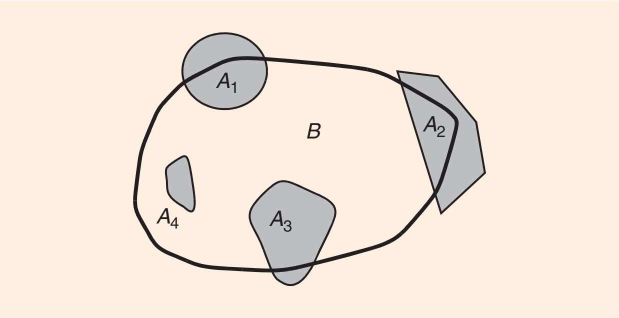
INVERSION OF THE INTERIOR HILBERT TRANSFORM

The problem of the interior finite Hilbert transform refers to the case $[h_L, h_R] \subset (f_L, f_R)$ where, as usual, the unknown function is zero outside $[f_L, f_R]$ and its Hilbert transform is known only on $[h_L, h_R]$. In this case we know that $f(x)$ cannot be stably reconstructed for $x \in (h_L, h_R)$, because such a result could be applied to the interior problem, and contradict the known nonuniqueness.

However, it has recently been shown [79]–[82] that if a small region of values of $f(x)$ are known inside the interior region, then uniqueness is restored and stable reconstruction can be achieved on the interior region.

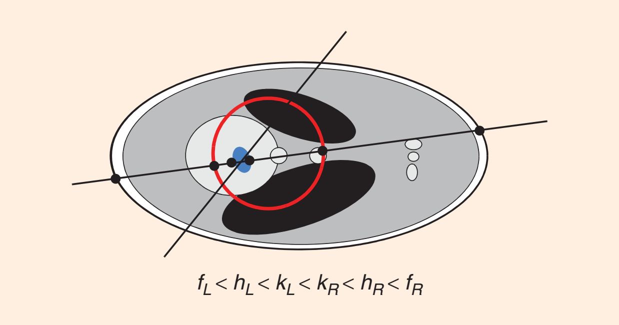
Figure 12 illustrates the situation in the image reconstruction context. The red circle indicates the FOV: only those lines crossing the red circle are measured by the scanner, and since the red circle doesn't meet the (known) boundary of the object, we are in the situation of the interior problem. In this case, region A = region C, and region B is minimal in the sense that it is not possible to remove (an open set of) measurement lines without reducing region A. Some information must be added to the problem to ensure unique reconstruction inside region A. In some practical imaging situations, it may be reasonable to assume a known value for the density at a certain location. If a small such region K of known values exists inside region A such as illustrated in Figure 12, then a new version of the finite Hilbert transform can be formulated that incorporates this new information.

Consider any line passing through region K (assumed to be within region A) as shown in Figure 12. Along this line, we identify the intervals $[f_L, f_R]$, $[h_L, h_R]$, and $[k_L, k_R]$, where as usual, $[f_L, f_R]$ is the known support of the object and $[h_L, h_R]$ is the known region of the Hilbert transform of the object along the line. Within the interval $[k_L, k_R]$ the values of the density function are known. The containment relation is $[k_L, k_R] \subset [h_L, h_R] \subset [f_L, f_R]$, and it was proved that the function could be stably reconstructed in (h_L, h_R). Applying this reasoning to all lines through region K establishes that all of region A can be stably reconstructed. However, similarly to the situation in the section "Inversion of the One-Sided Finite Hilbert Transform," no explicit inversion formula is known for this case of interior data with prior information.

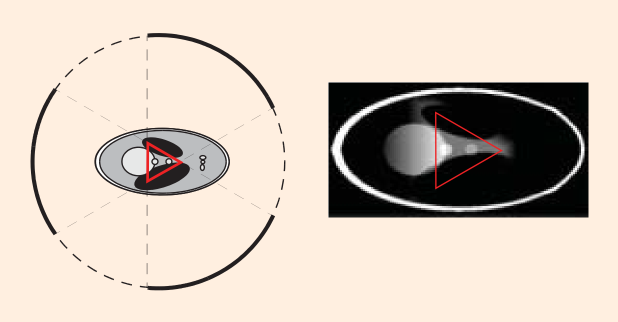


[FIG11] A general incomplete data problem. The object is convex, and in this example there are four separate connected components to (the completely measured) region C. These components must be convex. The three components of region A that touch the boundary (A_1, A_2, A_3) can be stably reconstructed, according to the one-sided Hilbert transform result [78], [56]. (Region A_2 can also be reconstructed using VFB or DBP-H methods.) For internal region A_4, it is unknown whether stable reconstruction is possible or not.

As a point of terminology, the "interior problem" refers to a particular image reconstruction geometry that is known to be mathematically unsolvable. Strictly speaking, it is not the interior problem that was 'solved' by adding new information; the supplementary information implied a new mathematical problem for which a stable solution exists. It is more convenient to describe this situation differently from the "interior problem with prior knowledge." We first note that knowing $f(x)$ for $x \in K$ is equivalent to $f(x) = 0$ for $x \in K$, because the known contributions of region K can be subtracted from the measurements. We can now regard region K as being outside the object, and the object boundary as including the hole



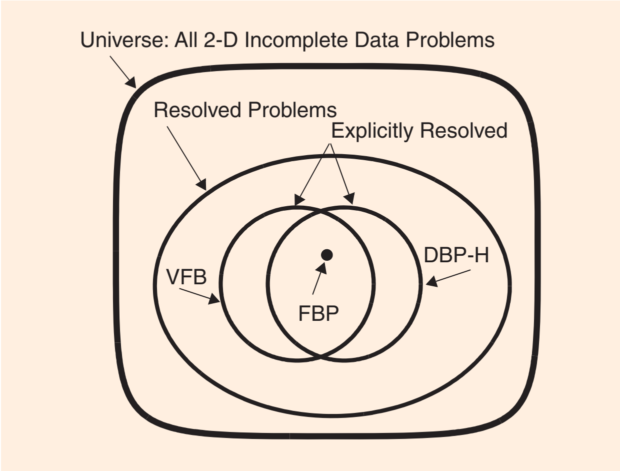
[FIG12] The interior problem with small a priori information. Measurements are only made along lines passing through the red FOV, resulting in the interior problem. However, if the density value is known a priori in a small region (region K, in blue), then the object density can be stably reconstructed inside the FOV.



[FIG13] An example where region A is internal to the object, yet can be reconstructed stably. Complete fanbeam projections are measured on a reduced scan consisting of three segments of 65° each. Region A is the red triangle and reconstruction was performed using the method of (16) and (17). For all other parts of the image, a unique solution exists but the inverse is not stable.

formed by region K. We refer to a convex object with a single such hole as a donut. The hole is not considered to be part of the donut.

It is now possible to combine the results on the donut problem with those of the one-sided Hilbert transform (see the section "Inversion of the One-Sided Finite Hilbert Transform") under a common framework. For convex or donuts-shaped objects (with known support in either case) the unknown density function can be stably reconstructed on each connected component of region A that extends to the object boundary. (Strictly speaking, there should be some of region C on the other side of the boundary, but in practice this must happen if region A touches the boundary. A technically precise statement is that $f(x)$ can be stably reconstructed (or is known to be zero) for x in any connected component of region C that is not entirely contained inside the convex- or donut-shaped object.)



[FIG14] Schematic description of the current state of the art in ROI reconstruction from incomplete data. The exterior box represents the set of all possible incomplete sinograms (including the trivial case of a sinogram whose set of missing measurements is empty). The single case of complete data is indicated with the point in the center, labeled "FBP."

In the situation of the interior problem, an alternative to adding information by providing values of $f(x)$ could be to provide more measurement lines. The diagram of Figure 13 has appeared in [61] and [70] (with a reconstructed example in [61]) and illustrates an imaging situation where region A is strictly internal to the object support, and yet stable reconstruction is possible using the Hilbert projection equality (see the section "The Parallel-Fanbeam Hilbert Projection Equality"). In this case, region B is not minimal, as there are many measurement lines that can be removed without affecting region A. This example shows that the VFB method may be useful to resolve other cases of an internal region A. Another such example appears in [65].

SUMMARY

A significant advance in 2-D image reconstruction theory took place at the turn of the century: accurate, robust ROI reconstruction from incomplete data was found to be possible despite intensive research suggesting the contrary during the late 1970s through the early 1990s. In this article, we have summarized the main advances that have occurred over the past eight years. The current situation for 2-D ROI reconstruction falls into three categories. The first consists of a collection of incomplete data problems for which explicit inversion formulas (and analytic reconstruction algorithms) can be applied. These are the cases that can be resolved using the VFB or DBP-H techniques described in the sections "The Parallel-Fanbeam Hilbert Projection Equality" and "Differentiated Backprojection with Hilbert Filtering." The second category comprises the incomplete data problems for which it is known that unique and stable reconstruction is possible but for which no explicit inversion formula (or direct analytic reconstruction algorithm) currently exists. The last category consists of those cases that are not resolved; incomplete data configurations for which it is not known which parts of the object, if any, can be reliably reconstructed. Figure 14 illustrates this situation in broad terms.

Twentieth century image-reconstruction theory does not contradict these new partial-data ROI reconstruction results. It was the conclusions extracted from that theory that were incorrect. The FBP formulation of Radon's inversion formula still stands and still requires complete sinograms for every possible ROI reconstruction. What is now understood is that the 2-D Radon transform operator has a special structure that admits multiple inversion formulas which use the redundant sinogram data in different ways to express the same unique inverse. It is these multiple nonequivalent reconstruction formulas that allow the phenomenon of partial reconstruction from partial data. The arguments for establishing instability of limited-angle reconstruction problems still hold in the ROI reconstruction arena and are the reason why valid ROI reconstruction can only exist in "region A" of the object. Truncated projections was possibly the area least thoroughly analyzed in the 20th century, the main result being the nonuniqueness of the interior problem. Even combined with the weight of other evidence, the conclusion that truncated projections could not be tolerated for accurate partial reconstruction was the least solid, and where current ROI reconstruction theory has penetrated the furthest. If the boundary of the object is not visible in any of the projection data (the interior problem), then indeed no accurate reconstruction can be achieved. Consequently all the ROI reconstruction examples given so far involve at least some projections seeing part of the object boundary. Moreover, for convex and donut-shaped objects, if part of the boundary is visible from all projection angles (i.e., is intersected by "region C") then ROI reconstruction is assured for all object points that form a "visible connection" to the boundary in all projections (i.e., reconstruction is assured in that connected component of region C). The unresolved situation is when no part of the boundary is visible from all projection angles, but a piece of boundary is sometimes visible and sometimes truncated in the projections (region A doesn't intersect the boundary). In a few such cases, such as that of Figure 13, stable reconstruction has been proved, but this case of "sometimes-visible boundaries" is currently where ROI reconstruction is the least understood.

The open challenges now are to resolve the ROI reconstruction problem for general incomplete data scenarios and to have explicit inversion formulas for all cases where ROI reconstruction is possible. In the coming years, one can anticipate new results that steadily build a complete understanding of 2-D ROI tomography and finally achieve the definitive theory for 2-D image reconstruction that was prematurely anticipated last century.

ACKNOWLEDGMENTS

The authors thank Dr. Catherine Mennessier of CPE/Lyon, France, for the computer simulation images of Figures 7 and 9, and for many helpful suggestions on the text. This work was partially supported by ANR grant NT05-1_45428 and the F.W.O. grant G.0569.08.

AUTHORS

Rolf Clackdoyle (rolf.clackdoyle@univ-st-etienne.fr) received his Ph.D. in mathematics from Dalhousie University, Canada, in 1989. From 1992 to 2004, he was an assistant professor, associate professor, and professor of radiology at the University of Utah and a member of the Medical Imaging Research Laboratory (now the Utah Center for Advanced Imaging Research). Since 2005, he has been with the Centre National de la Recherche Scientifique (CNRS), where he holds the rank of directeur de recherche. He is a member of the CNRS–Saint Etienne University mixed research unit, the Laboratoire Hubert Curien. His research interests are in image reconstruction from projections using analytic methods, mainly directed at medical imaging applications.

Michel Defrise (mdefrise@vub.ac.be) received the Ph.D. degree in theoretical physics from the University of Brussels in 1981. He was a visiting professor in the Department of Radiology of the University of Geneva in 1992 and 1993. He is currently a research professor in the Department of Nuclear Medicine at the VUB University Hospital in Brussels. He has participated actively in the advancement of 3-D PET methodology. His current research interests include 3-D image reconstruction in nuclear medicine (PET and SPECT) and in CT.

REFERENCES

[1] G. T. Herman, *Image Reconstruction from Projections*. New York: Academic, 1980.

[2] S. R. Deans, *The Radon Transform and Some of Its Applications*. New York: Wiley, 1983.

[3] F. Natterer, *The Mathematics of Computerized Tomography*. New York: Wiley, 1986.

[4] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. New York: IEEE Press, 1988.

[5] R. A. Brooks and G. Di Chiro, "Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging," *Phys. Med. Biol.*, vol. 21, no. 5, pp. 689–732, Sept. 1976.

[6] H. J. Scudder, "Introduction to computer aided tomography," *Proc IEEE*, vol. 66, no. 6, pp. 628–637, 1978.

[7] C. L. Byrne, *Applied Iterative Methods*. Natick, MA: Peters, 2008.

[8] J. Qi and R. M. Leahy, "Iterative reconstruction techniques in emission computed tomography," *Phys. Med. Biol.*, vol. 51, no. 15, pp. R541–R578, 2006.

[9] J. A. Fessler, "Statistical image reconstruction methods for transmission tomography," in *Handbook of Medical Imaging, vol. 2, Medical Image Processing and Analysis*, M. Sonka and J. M. Fitzpatrick, Eds. Bellingham: SPIE, 2000, pp. 1–70.

[10] D. L. Parker, "Optimal short scan convolution reconstruction for fan-beam CT," *Med. Phys.*, vol. 9, no. 2, pp. 254–257, 1982.

[11] R. M. Perry, "On reconstruction of a function on the exterior of a disc from its Radon transform," *J. Math. Anal. Appl.*, vol. 59, no. 2, pp. 324–341, 1977.

[12] B. E. Oppenheim, "Reconstruction tomography from incomplete projections," in *Reconstruction Tomography in Diagnostic Radiology and Nuclear Medicine*, M. M. Ter-Pogossian, Ed. Baltimore: Univ. Park Press, 1977, pp. 155–183.

[13] R. M. Lewitt and R. H. T. Bates, "Image reconstruction from projections. III. Projection completion methods (theory)," *Optik*, vol. 50, no. 3, pp. 180–204, 1978.

[14] R. M. Lewitt and R. H. T. Bates, "Image reconstruction from projections. IV. Projection completion methods (computational examples)," *Optik*, vol. 50, no. 3, pp. 269–278, 1978.

[15] R. M. Lewitt, "Processing of incomplete measurement data in computed tomography," *Med. Phys.*, vol. 6, no. 5, pp. 412–417, 1979.

[16] O. Nalcioglu, P. V. Sankar, and J. Sklansky, "Region-of-interest X-ray tomography (ROIT)," *SPIE*, vol. 206, pp. 98–102, 1979.

[17] O. Nalcioglu, Z. H. Cho, and R. Y. Lou, "Limited field of view reconstruction in computerized tomography," *IEEE Trans. Nucl. Sci.*, vol. 26, no. 1, pp. 546–551, 1979.

[18] W. Wagner, "Reconstruction from restricted region scan data—New means to reduce the patient dose," *IEEE Trans. Nucl. Sci.*, vol. no. 2, NS-26, pp. 2866–2869, 1979.

[19] J. C. Gore and S. Leeman, "The reconstruction of objects from incomplete projections," *Phys. Med. Biol.*, vol. 25, no. 1, pp. 129–136, 1980.

[20] A. Lent and H. Tuy, "An iterative method for the extrapolation of band limited functions," *J. Math. Anal. Appl.*, vol. 83, no. 2, pp. 554–565, 1981.

[21] K. C. Tam and V. Perez-Mendes, "Tomographical imaging with limited angle input," *J. Opt. Soc. Amer.*, vol. 71, no. 5, pp. 582–592, 1981.

[22] T. Sato, S. J. Norton, M. Linzer, O. Ikeda, and M. Hirama, "Tomographic imaging reconstruction from limited projections using iterative revisions in image and transform space," *J. Opt. Soc. Amer.*, vol. 71, pp. 582–592, May 1981.

[23] P. V. Sankar, O. Nalcioglu, and J. Sklansky, "Undersampling errors in region-of-interest tomography," *IEEE Trans. Med. Imaging*, vol. MI-1, no. 3, pp. 168–173, 1982.

[24] T. Inouye, "Image reconstruction with limited view angle projections," in *Proc. Int. Workshop Physics and Engineering in Medical Imaging*, Pacific Grove, CA, Mar. 1982, pp. 165–168.

[25] M. Nassi, W. R. Brody, B. P. Medoff, and A. Macovski, "Iterative reconstruction-reprojection: An algorithm for limited data cardiac computed tomography," *IEEE Trans. Biomed. Eng.*, vol. BME-29, pp. 331–341, May 1982.

[26] J. S. Choi, K. Ogawa, M. Nakajima, and S. Yuta, "A reconstruction algorithm of body section s with opaque obstructions," *IEEE Trans. Sonics Ultrason.*, vol. SU-29, pp. 143–150, May 1982.

[27] H. K. Tuy, "An algorithm for incomplete range of views reconstruction," in *Tech. Dig. Topical Meeting Signal Recovery and Synthesis with Incomplete Information and Partial Constraints*, NV, Jan. 1983, pp. FA1-1, FA1-4.

[28] M. Davison, "The ill-conditioned nature of the limited angle tomography problem," *SIAM J. Appl. Math.*, vol. 43, no. 2, pp. 428–448, 1983.

[29] B. P. Medoff, W. R. Brody, M. Nassi, and A. Macovski, "Iterative convolution backprojection algorithms for image reconstruction from limited data," *J. Opt. Soc. Amer.*, vol. 73, pp. 1493–1500, Nov. 1983.

[30] K. M. Hanson and G. W. Wecksung, "Bayesian approach to limited angle reconstruction in computed tomography," *J. Opt. Soc. Amer.*, vol. 73, pp. 1501–1509, Nov. 1983.

[31] M. I. Sezan and H. Stark, "Tomographic image reconstruction from incomplete view data by convex projections and direct Fourier inversion," *IEEE Trans. Med. Imaging*, vol. MI-3, no. 2, pp. 91–98, 1984.

[32] K. Ogawa, M. Nakajima, and S. Yuta, "A reconstruction algorithm from truncated projections," *IEEE Trans. Med. Imaging*, vol. MI-3, no. 1, pp. 34–40, 1984.

[33] J. H. Kim, K. Y. Kwak, S. B. Park, and Z. H. Cho, "Projection space iteration reconstruction-reprojection," *IEEE Trans. Med. Imaging*, vol. MI-4, pp. 139–143, Sept. 1985.

[34] N. Srinivasa, V. Krishnan, K. R. Ramakrishnan, and K. Rajgopal, "Image reconstruction from truncated projections: A linear prediction approach," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, pp. 1733–1736.

[35] A. Louis, "Incomplete data problems in X-ray computerized tomography 1. Singular value decomposition of the limited angle transform," *Numer. Math.*, vol. 48, no. 3, pp. 251–262, 1986.

[36] W. Madych and S. Nelson, "Reconstruction from restricted Radon transform data: resolution and ill-conditionedness," *SIAM J. Math. Anal.*, vol. 17, no. 6, pp. 1447–1453, 1986.

[37] J. A. Reeds and L. A. Shepp, "Limited angle reconstruction in tomography via squashing," *IEEE Trans. Med. Imaging*, vol. MI-6, pp. 89–97, June 1987.

[38] B. P. Medoff, "Image reconstruction from limited data: Theory and applications in computerized tomography," in *Image Recovery: Theory and Applications*, H. Stark, Ed. New York: Academic, 1987, pp. 321–368.

[39] E. T. Quinto, "Tomographic reconstructions from incomplete data—numerical inversion of the exterior Radon transform," *Inverse Probl.*, vol. 4, no. 3, pp. 867–876, 1988.

[40] H. Peng and H. Stark, "One-step image reconstruction from incomplete data incomputer tomography," *IEEE Trans. Med. Imaging*, vol. 8, no. 1, pp. 16–31, 1989.

[41] P. Oskoui and H. Stark, "A comparative study of three reconstruction methods for a limited-view computer tomography problem," *IEEE Trans. Med. Imaging*, vol. 8, no. 1, pp. 43–49, 1989.

[42] A. Louis and A. Rieder, "Incomplete data problems in X-ray computerized tomography II. Truncated projections and region-of-interest tomography," *Numer. Math.*, vol. 56, no. 4, pp. 371–383, 1989.

[43] K. C. Tam, J. W. Eberhard, and K. W. Mitchell, "Incomplete-data CT image reconstructions in industrial applications," *IEEE Trans. Nucl. Sci.*, vol. 37, no. 3, pp. 1490–1499, 1990.

[44] H. Kudo and T. Saito, "Sinogram recovery with the method of convex projection for limited-data reconstruction in computed tomography," *J. Opt. Soc. Amer.*, vol. 8, no. 7, pp. 1148–1160, 1991.

[45] P. Maass, "The interior Radon transform," *SIAM J. Appl. Math.*, vol. 52, no. 3, pp. 710–724, 1992.

[46] H. Peng and H. Stark, "Image recovery in computer tomography from partial fan-beam data by convex projections," *IEEE Trans. Med. Imaging*, vol. 11, no. 4, pp. 470–478, Dec. 1992.

[47] E. T. Quinto, "Singularities of the X-ray transform and limited data tomography in R^2 and R^3," *SIAM J. Math. Anal.*, vol. 24, no. 5, pp. 1215–1225, 1993.

[48] E. T. Quinto, "Exterior and limited angle tomography in nondestructive evaluation," *Inverse Probl.*, vol. 14, no. 2, pp. 339–353, 1998.

[49] B. Ohnesorge, T. Flohr, and K. Schwarz, "Efficient correction for CT image artifacts caused by objects extending outside the scanner field of view," *Med. Phys.*, vol. 27, no. 1, pp. 39–46, 2000.

[50] J. Hsieh, E. Chao, J. Thibault, B. Grekowicz, A. Horst, S. McOlash, and T. J. Myers, "A novel reconstruction algorithm to extend the CT scanner field-of-view," *Med. Phys.*, vol. 31, no. 9, pp. 2385–2391, 2004.

[51] B. Zhang and G. L. Zeng, "Two-dimensional iterative region-of-interest (ROI) reconstruction from truncated projection data," *Med. Phys.*, vol. 34, no. 3, pp. 935–944, 2007.

[52] K. T. Smith, D. C. Solmon, S. L. Wagner, and C. Hamaker, " Mathematical aspects of divergent beam radiography," *Proc. Nat. Acad. Sci. USA*, vol. 75, no. 5, pp. 2055–2058, May 1978.

[53] C. Hamaker, K. T. Smith, D. C. Solmon, and S. L. Wagner, "The divergent beam X-ray transform," *Rocky Mountain J. Math.*, vol. 10, no. 1, pp. 253–283, 1980.

[54] D. Slepian, H. O. Pollak, and H. J. Landau, "Prolate spheroidal wave functions I, 11," *Bell Syst. Tech. J.*, vol. 40, no. 1, pp. 43–84, 1961.

[55] M. Bertero, G. A. Viano, and C. DeMol, "Resolution beyond the Rayleigh limit for regularizing object restoration," *Opt. Acta*, vol. 27, no. 3, pp. 307–320, 1980.

[56] R. Clackdoyle, M. Defrise, F. Noo, and H. Kudo, "Two-dimensional region-of-interest tomography," in *Oberwolfach Reports*, vol. 3, no. 3, G.-M. Greuel and S. Klaus, Eds. Zurich, Switzerland: European Mathematical Society Publishing House, 2006, pp. 2070–2073.

[57] A. Faridani, E. L. Ritman, and K. T. Smith, "Local tomography," *SIAM J. Appl. Math.*, vol. 52, no. 1, pp. 459–484, 1992.

[58] D. V. Finch, "Cone beam reconstruction with sources on a curve," *SIAM J. Appl. Math.*, vol. 45, no. 4, pp. 665–673, 1985.

[59] H. K. Tuy, "An inversion formula for cone-beam reconstruction," *SIAM J. Appl. Math.*, vol. 43, no. 3, pp. 546–552, 1983.

[60] R. Clackdoyle, F. Noo, J. Guo, and J. A. Roberts, "Quantitative reconstruction from truncated projections in classical tomography," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 5, pp. 2570–2578, 2004.

[61] F. Noo, M. Defrise, R. Clackdoyle, and H. Kudo, "Image reconstruction from fan-beam projections on less than a short-scan," *Phys. Med. Biol.*, vol. 47, no. 14, pp. 2525–2546, 2002.

[62] L. A. Shepp and B. F. Logan, "The Fourier reconstruction of a head section ," *IEEE Trans. Nucl. Sci.*, vol. NS-21, pp. 21–43, June 1974.

[63] H. Kudo, F. Noo, M. Defrise, and R. Clackdoyle, "New super-short-scan algorithms for fan-beam and conebeam reconstruction," in *Conf. Rec. 2002 IEEE Nuclear Science Symp. and Medical Imaging Conf.*, Norfolk, VA, IEEE Service Center, 2003, vol. 2, pp. 902–906.

[64] R. Clackdoyle, F. Noo, M. S. Ould Mohamed, and C. Mennessier, "Filtered-backprojection reconstruction formula for 2D tomography with bilateral truncation," in *Conf. Rec. 2006 IEEE Nuclear Science Symp. Medical Imaging Conf.*, San Diego, 2006, vol. 5, pp. 2895–2899.

[65] M. S. Ould Mohamed, R. Clackdoyle, and C. Mennessier, "Region of interest reconstruction from truncated data by combined classical filtered backprojection and virtual fanbeam reconstruction," in *Conf. Rec. 2008 IEEE Nuclear Science Symp. Medical Imaging Conf.*, Dresden, 2008, pp. 4178–4181.

[66] R. Clackdoyle and F. Noo, "A large class of inversion formulae for the 2-D Radon transform of functions of compact support," *Inverse Probl.*, vol. 20, no. 4, pp. 1281–1291, 2004.

[67] R. Clack, "Towards a complete description of 3D filtered backprojection," *Phys. Med. Biol.*, vol. 37, no. 3, pp. 645–660, 1992.

[68] I. M. Gelfand and M. I. Graev, "Crofton's function and inversion formulas in real integral geometry," *Funct. Anal. Applicat.*, vol. 25, no. 1, pp. 1–5, 1991.

[69] D. V. Finch, *Mathematisches Forschungsinstitut Oberwolfach (private communication)*, Aug. 2002.

[70] F. Noo, R. Clackdoyle, and J. D. Pack, "A two-step Hilbert transform method for 2-D image reconstruction," *Phys. Med. Biol.*, vol. 49, no. 17, pp. 3903–3923, 2004.

[71] T. Zhuang, S. Leng, B. E. Nett, and G.-H. Chen, "Fan-beam and cone-beam image reconstruction via filtering the backprojection image of differentiated projection data," *Phys. Med. Biol.*, vol. 49, no. 24, pp. 5489–5503, 2004.

[72] Y. Zou, X. Pan, and E. Y. Sidky, "Image reconstruction in regions-of-interest from truncated projections in a reduced fan-beam scan," *Phys. Med. Biol.*, vol. 50, no. 1, pp. 13–28, 2005.

[73] Y. Zou and X. Pan, "An extended data function and its generalized backprojection for image reconstruction in helical cone-beam CT," *Phys. Med. Biol.*, vol. 49, no. 22, 2004, p. N383.

[74] Y. Zou and X. Pan, "Exact image reconstruction on PI-lines from minimum data in helical cone-beam CT," *Phys. Med. Biol.*, vol. 49, no. 6, pp. 941–959, 2004.

[75] J. D. Pack, F. Noo, and R. Clackdoyle, "Cone-beam reconstruction using the backprojection of locally filtered projections," *IEEE Trans. Med. Imaging*, vol. 24, no. 1, pp. 70–85, 2005.

[76] Y. Ye, S. Zhao, H. Yu, and G. Wang, "A general exact reconstruction for cone-beam CT via backprojection-filtration," *IEEE Trans. Med. Imaging*, vol. 24, no. 9, pp. 1190–1198, 2005.

[77] F. G. Tricomi, *Integral Equations*. New York: Dover, 1957.

[78] M. Defrise, F. Noo, R. Clackdoyle, and H. Kudo, "Truncated Hilbert transform and image reconstruction from limited tomographic data," *Inverse Probl.*, vol. 22, no. 3, pp. 1037–1053, 2006.

[79] Y. Ye, H. Yu, Y. Wei, and G. Wang, "A general local reconstruction approach based on a truncated Hilbert transform," *Int. J. Biomed. Imaging*, 2007.

[80] Y. Ye, H. Yu, and G. Wang, "Exact interior reconstruction with cone-beam CT," *Int. J. Biomed. Imaging*, 2007.

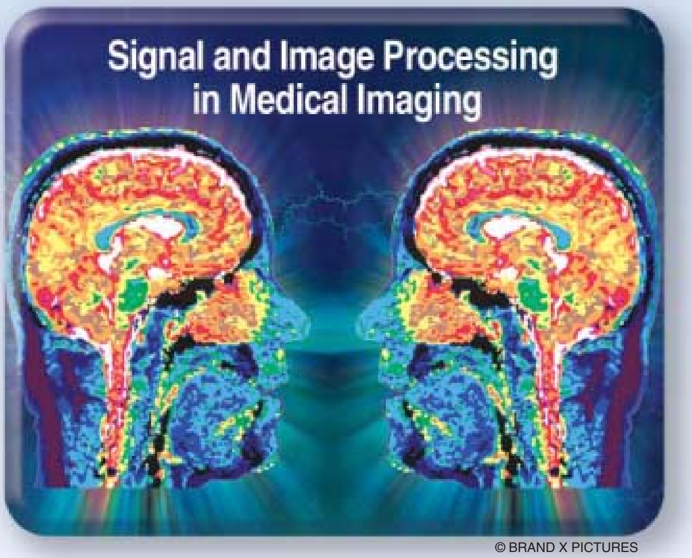
[81] M. Courdurier, F. Noo, M. Defrise, and H. Kudo, "Solving the interior problem of computed tomography using a priori knowledge," *Inverse Probl.*, vol. 24, no. 6, p. 065001, 2008.

[82] H. Kudo, M. Courdurier, F. Noo, and M. Defrise, "Tiny a priori knowledge solves the interior problem in computed tomography," *Phys. Med. Biol.*, vol. 53, no. 9, pp. 2207–2231, 2008.

[Jeffrey A. Fessler]

Model-Based Image Reconstruction for MRI

[A review of the use of iterative algorithms]



© BRAND X PICTURES

Magnetic resonance imaging (MRI) is a sophisticated and versatile medical imaging modality. Traditionally, MR images are reconstructed from the raw measurements by a simple inverse two-dimensional (2-D) or three-dimensional (3-D) fast Fourier transform (FFT). However, there are a growing number of MRI applications where a simple inverse FFT is inadequate, e.g., due to non-Cartesian sampling patterns, non-Fourier physical effects, nonlinear magnetic fields, or deliberate under-sampling to reduce scan times. Such considerations have led to increasing interest in methods for model-based image reconstruction in MRI.

INTRODUCTION

The inverse FFT has served the MR community very well as the conventional image reconstruction method for k-space data with full Cartesian sampling. And for well sampled non-Cartesian data, the gridding method with appropriate density compensation factors [1] is fast and effective. But when only under-sampled data is available, or when non-Fourier physical effects like field inhomogeneity are important, then gridding/FFT methods for image reconstruction are suboptimal, and iterative algorithms based on appropriate models can improve image quality, at the price of increased computation. This article reviews the use of iterative algorithms for model-based MR image reconstruction. The references give pointers to some recent work but are by no means a comprehensive survey. To see more citations, visit http://www.eecs.umich.edu/~fessler/.

MRI BACKGROUND

Any signal processing method aimed at forming images from measurement devices such as MRI scanners must consider the relevant physics. A survey in *IEEE Signal Processing Magazine* [2] and a book written from a signal processing perspective [3] have described MRI physics well. Here we review the physics in a somewhat unconventional way that facilitates describing some of the "non-Fourier" aspects of MRI.

MRI PHYSICS

Standard MRI scanners use a large static magnetic field

$$\vec{B}_0(\vec{r}\,) = B_0(\vec{r}\,)\vec{k} \qquad (1)$$

Digital Object Identifier 10.1109/MSP.2010.936726

to induce a net magnetization $\vec{M} = M_x\vec{i} + M_y\vec{j} + M_z\vec{k}$ at each point in space in the body being imaged, where $\vec{i}$, $\vec{j}$, and $\vec{k}$ denote the unit vectors along the x, y, and z axes, respectively, and $\vec{r} = (x, y, z)$ denotes 3-D spatial coordinates. Ideally, the static field strength $B_0(\vec{r})$ would be spatially uniform, i.e., a single constant B_0. In practice, it is never perfectly uniform, due to the unavoidable nonuniformities of all practical coil designs and due to the field strength variations that are induced by the nonuniform magnetic susceptibilities of different tissue types. The electron distributions in different molecules also influence the local magnetic environment experienced by an atom's nucleus, called chemical shift. Some types of MRI scans are robust to such spatial variations of B_0; others are sensitive to nonuniformities, necessitating correction methods.

At equilibrium (which is established within a few seconds for a stationary object), the magnetization $\vec{M}$ is aligned with the applied static field and its magnitude is proportional to the product of $B_0(\vec{r})$ and the object-dependent local density of (predominately) hydrogen protons or "spins." This proton density alone is of only modest interest in MRI; in practice one applies time-varying magnetic fields $\vec{B}(\vec{r}, t)$ that induce time-varying changes in the magnetization

$$\vec{M}(\vec{r}, t) = M_x(\vec{r}, t)\vec{i} + M_y(\vec{r}, t)\vec{j} + M_z(\vec{r}, t)\vec{k}. \qquad (2)$$

These changes depend on time constants (tissue-dependent relaxation parameters) and other factors, and the goal in MRI is to form images of aspects of this magnetization. By manipulating the applied field $\vec{B}_0(\vec{r}, t)$ appropriately, sometimes in conjunction with injected or inhaled contrast agents, one can examine a multitude of different tissue properties.

An MRI scan consists of one or more alternations between two stages: excitation and readout. During the excitation stage, the applied magnetic field $\vec{B}(\vec{r}, t)$ is designed to tip the magnetization vectors $\vec{M}$ within some slice or slab away from equilibrium, so that they have a component in the transverse plane, i.e., the (x, y) plane. It is convenient to represent this transverse component mathematically using a complex function defined as follows:

$$\mathrm{M}(\vec{r}, t) \triangleq M_x(\vec{r}, t) + i\,M_y(\vec{r}, t), \qquad (3)$$

where $i \triangleq \sqrt{-1}$. Note that the field components M_x and M_y are real physical quantities; the "transverse magnetization" $\mathrm{M}(\vec{r}, t)$ is complex solely by definition. The excitation process can be quite complicated to model and is beyond the scope of this article. See [2] for an introduction to the role that signal processing plays in the design of excitation pulses and [4] for some recent model-based RF pulse design methods.

During the readout stage, the applied field $\vec{B}(\vec{r}, t)$ is manipulated in ways that help elucidate the transverse magnetization $\mathrm{M}(\vec{r}, t)$. For image reconstruction, it is essential to model the effects of the applied field on the transverse magnetization. The precise relationship is governed by the Bloch equation [2]. For most image reconstruction purposes, it suffices to consider just two aspects of the full relationship: precession and transverse relaxation. The most important equation in MRI is the Larmor relation: $\omega = \gamma|\vec{B}|$, which states that the magnetization precesses (around the axis of the applied field) at a frequency ω that is proportional to the magnitude of the applied field. The constant of proportionality γ is called the gyromagnetic ratio and is about 42.6 MHz/T for hydrogen protons. During a readout, only the longitudinal component of $\vec{B}$ is varied usually, i.e.,

$$\vec{B}(\vec{r}, t) = B_z(\vec{r}, t)\vec{k}, \qquad (4)$$

so the magnetization precesses around $\vec{k}$, i.e., within the transverse plane. This property is why the complex representation (3) is convenient, because precession can be expressed using a complex phase in this form. In general, the applied longitudinal field strength $B_z(\vec{r}, t)$ varies both spatially and temporally, so the Larmor relationship describes the instantaneous frequency at a given spatial location

$$\omega(\vec{r}, t) = \gamma\, B_z(\vec{r}, t). \qquad (5)$$

Without loss of generality, let $t = 0$ be the time when the excitation pulse is completed, and consider some time point $t > 0$ during the readout. The precession of the transverse magnetization between time zero and time t corresponds to a net phase that is the integral of the instantaneous frequency (5), i.e., ideally we would have

$$\mathrm{M}(\vec{r}, t) = \mathrm{M}(\vec{r}, 0)\exp\left(-i\int_0^t \omega(\vec{r}, t')\,\mathrm{d}t'\right).$$

In practice, microscopic variations in the magnetic field cause the spins within a given voxel to become out of phase over time. So the transverse magnetization vector's magnitude decreases approximately exponentially with a time constant T_2^*. Accounting for this decay, an accurate model for the temporal evolution of the transverse magnetization during a readout is

$$\mathrm{M}(\vec{r}, t) = f(\vec{r})e^{-t/T_2^*(\vec{r})}\exp\left(-i\gamma\int_0^t B_z(\vec{r}, t')\,\mathrm{d}t'\right), \qquad (6)$$

where $f(\vec{r}) \triangleq \mathrm{M}(\vec{r}, 0)$ denotes the object's transverse magnetization immediately after excitation. A typical goal in MRI is to form an image of $f(\vec{r})$. The properties of $f(\vec{r})$ depend not only on spin density, but also on the type of excitation used. Note that for simplicity of exposition, we focus here on the case where the object is static so that $f(\vec{r})$ is not a function of time t. Generalizations to dynamic imaging are very active research areas in MR image reconstruction.

The relaxation factor T_2^* varies spatially, and often is on the order of 10 ms. This relatively rapid decay is a significant limitation in MRI. If T_2^* were longer, then a signal excitation stage followed by a (lengthy) readout stage could be sufficient to form a high-resolution image of $f(\vec{r})$. In practice, the rapid decay limits how much spatial information can be recorded in a single readout stage, so such "single shot" imaging, such as echo-planar imaging (EPI) [5], provides only modest spatial resolution. Therefore, high-resolution imaging uses multiple alternations between excitation stages and readout stages, each with different variations of the applied field $B_z(\vec{r}, t)$.

[NOTIONS OF SPARSITY HAVE DEEP ROOTS IN STATISTICAL SIGNAL PROCESSING.]

DATA ACQUISITION: THE MR SIGNAL

By Faraday's law, the time-varying magnetization $\mathrm{M}(\vec{r}, t)$ will induce an electromotive force (emf) in a nearby coil. The emf will be proportional to the volume integral of the time derivative of the magnetization $\mathrm{M}(\vec{r}, t)$ multiplied by the coil response pattern $c(\vec{r})$. The resulting electrical potential $v(t)$ across the receive coil is

$$v(t) = \mathrm{real}\left(\int c(\vec{r}) \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{M}(\vec{r}, t) \mathrm{d}\vec{r}\right), \tag{7}$$

where real(·) denotes the real part of a complex number. The coil response $c(\vec{r})$ generally decreases with distance from the coil. If uncorrected, this nonuniformity causes spatial variations in signal strength that can be a challenge for image processing methods like segmentation algorithms. Numerous correction methods have been developed.

Because the time constant T_2^* is on the order of milliseconds whereas the phase variations in (6) are many MHz, it is very reasonable to use a narrow-band approximation when evaluating the time derivative of $\mathrm{M}(\vec{r}, t)$ as needed in (7). The time derivative of a narrow-band signal is well approximated by a constant scaling factor $\mathrm{d}/\mathrm{d}t\mathrm{M}(\vec{r}, t) \approx c_0\, \mathrm{M}(\vec{r}, t)$. We absorb this constant into the coil response pattern and rewrite (7) as

$$v(t) = \mathrm{real}\left(\int c(\vec{r})\, \mathrm{M}(\vec{r}, t)\, \mathrm{d}\vec{r}\right). \tag{8}$$

The receive coil's signal is amplified and demodulated using some center frequency ω_0. Ideally, one would use $\omega_0 = \gamma B_0$ if the static magnetic field had uniform strength B_0. Usually quadrature demodulation is used, yielding separate in-phase $I(t)$ and quadrature $Q(t)$ baseband signals. In the literature, the demodulated "MR signal" $s(t)$ is defined (implicitly) as

$$s(t) \triangleq I(t) + iQ(t) = \mathrm{lowpass}(\mathrm{e}^{i\omega_0 t} v(t)) = \mathrm{e}^{i\omega_0 t} \int c(\vec{r}) \mathrm{M}(\vec{r}, t) \mathrm{d}\vec{r}, \tag{9}$$

where the low-pass operation selects the baseband component of the demodulated signal. This complex analog signal is just a mathematical definition; in practice, the $I(t)$ and $Q(t)$ signals are each sampled and digitized yielding two digital signals. (One can use two separate analog-to-digital (A/D) converters, or a single A/D converter running at twice the normal rate to avoid I/Q imbalance.) Digitally, these two signals can be combined and stored as complex values, i.e., we record samples

$$I(m\Delta_{\mathrm{T}}) + iQ(m\Delta_{\mathrm{T}}), \quad m = 1, \ldots, n_d,$$

where Δ_{T} denotes the sampling rate (typically around 1 μs) and n_{d} denotes the number of recorded samples, typically 64–512 for a given readout stage. Again, the physical quantities are real, but complex quantities are defined in terms of those physical quantities for convenience. (In some systems, digital demodulation is used, but the modeling remains identical.)

SIGNAL MODEL

To improve signal-to-noise ratio and reduce acquisition times, the use of multiple receive coils has become increasingly popular in MRI. Although originally called phased array imaging [6], a term that resonates with other signal processing applications involving multiple receivers, today the use of multiple receive coils in MRI is usually called parallel imaging [7].

Let $c_l(\vec{r})$ denote the sensitivity (response pattern) of the lth coil, for $l = 1, \ldots, L$, where L denotes the number of coils. Let $s_l(t)$ denote the demodulated "MR signal" associated with the lth coil, defined as in (9). Substituting (6) into (9) and simplifying yields the following general forward model for the MR signal associated with the lth coil

$$s_l(t) = \int c_l(\vec{r}) f(\vec{r}) \mathrm{e}^{-t/T_2^*(\vec{r})} \mathrm{e}^{-i\phi(\vec{r}, t)}\, \mathrm{d}\vec{r}, \tag{10}$$

where the space- and time-varying phase is

$$\phi(\vec{r}, t) \triangleq \int_0^t (\gamma B_z(\vec{r}, t') - \omega_0) \mathrm{d}t'. \tag{11}$$

In practice, multiple such signals are recorded, one for each excitation/readout pair ("shots"). For simplicity of notation, we consider "single shot" imaging; the extension to multiple shots is conceptually straightforward but notationally cumbersome. Note that the phase variations (11) are common to all receive coils; only the coil response patterns $\{c_l(\vec{r})\}$ differ between coils.

MEASUREMENT MODEL

The recorded measurements in a MR scan consist of noisy samples of the MR signal (10)

$$y_{li} = s_l(t_i) + \varepsilon_{li}, \quad i = 1, \ldots, n_d, \quad l = 1, \ldots, L, \tag{12}$$

where y_{li} denotes the ith sample of the lth coil's signal at time t_i and n_d denotes the number of time samples. Usually the t_i values are equally spaced, and often there are one or more time values where the signal is particularly strong due to alignment of the magnetization's phases; these values are called echo times. The measurement errors ε_{li} are very well modeled by additive, complex, zero-mean, temporally white Gaussian noise [8]. However, there can be coupling of the noise values between different coils for the same time points, i.e.,

$$\mathrm{Cov}\{\varepsilon_{li}, \varepsilon_{kj}\} = \Sigma_{lk}\delta[i-j], \tag{13}$$

where δ denotes the Kronecker impulse, and the $L \times L$ matrix Σ characterizes the noise covariance between coils [7].

MODEL-BASED METHODS THAT ACCOUNT FOR THOSE [PHYSICAL] EFFECTS ARE PROVING TO BE BENEFICIAL FOR IMPROVING IMAGE QUALITY.

LINEAR RECONSTRUCTION PROBLEM

Using the measurement model (12) and the signal model (10), the "typical" image reconstruction problem in MRI is to estimate the object $f(\vec{r})$ from the measurement vector $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L)$, where $\boldsymbol{y}_l = (y_{l1}, \ldots, y_{l,n_d})$. (All vectors are column vectors here.) We first consider model-based image reconstruction for this "basic" linear formulation. Because parallel imaging is of considerable interest, we continue to consider the general case of L receive coils. A standard single receive coil is a simple special case.

This is an ill-posed problem because the given measurements $\boldsymbol{y}$ are discrete whereas the object $f(\vec{r})$ is an unknown continuous-space function. To facilitate parametric estimation, we approximate the object $f(\vec{r})$ using a "finite series expansion" as follows:

$$f(\vec{r}) = \sum_{j=1}^{N} f_j b(\vec{r} - \vec{r}_j), \tag{14}$$

where $b(\cdot)$ denotes the object basis function, $\vec{r}_j$ denotes the center of the jth translated basis function, and N is the number of parameters. Such approximations are classic in the tomographic image reconstruction literature [9] and are slowly taking root in the MR community. Minimum $\mathcal{L}_2$ norm methods can postpone the discretization (14) until the final step of displaying the image, but it is unclear if this approach provides image quality benefits that outweigh its computational requirements. For simplicity, hereafter we use rect basis functions $b(\vec{r}) = \mathrm{rect}(\vec{r}/\Delta)$, i.e., square pixels of dimension Δ, so N is the number of pixels, or voxels in 3-D scans. Many other possible basis function choices can be considered, all of which are imperfect because the true object never satisfies the parametric model (14) exactly. Nevertheless simple basis functions can provide useful approximations.

Substituting the basis expansion (14) into the signal model (10) and simplifying leads to the discrete forward model

$$s_l(t_i) = \sum_{j=1}^{N} a_{lij} f_j, \tag{15}$$

where the elements $\{a_{lij}\}$ of the system matrix $\boldsymbol{A}_l$ associated with the lth coil are given by

$$a_{lij} = \int b(\vec{r} - \vec{r}_j) c_l(\vec{r}) \mathrm{e}^{-t_i/T_2^*(\vec{r})} \mathrm{e}^{-i\phi(\vec{r},\, t_i)}\, \mathrm{d}\vec{r}. \tag{16}$$

In practice the basis functions are usually highly localized (e.g., voxels), so "center of voxel" approximations like the following are nearly always used, often implicitly

$$a_{lij} \approx c_l(\vec{r}_j) \mathrm{e}^{-t_i/T_2^*(\vec{r}_j)} \mathrm{e}^{-i\phi(\vec{r}_j,\, t_i)}. \tag{17}$$

For exceptions, see [10].

Typically the decay due to T_2^* is ignored, or it is assumed implicitly that the total readout time $t_{n_d} - t_1$ is small relative to T_2^* in which case one can make the approximation $\mathrm{e}^{-t_i/T_2^*(\vec{r})} \approx \mathrm{e}^{-t_1/T_2^*(\vec{r})}$. Under this approximation, we can absorb the T_2^*-weighting effect of $\mathrm{e}^{-t_1/T_2^*(\vec{r})}$ into the unknown image $f(\vec{r})$.

Combining (12) and (15) in matrix-vector form yields

$$\boldsymbol{y}_l = \boldsymbol{A}_l \boldsymbol{f} + \boldsymbol{\varepsilon}_l,$$

where $\boldsymbol{f} = (f_1, \ldots, f_N)$ is the vector of parameters (pixel values) that we wish to estimate from the data $\boldsymbol{y}$. Stacking up all L measurement vectors as $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L)$ and defining the $(n_d L) \times N$ system matrix $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L)$ yields the linear model

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{\varepsilon}. \tag{18}$$

At first glance this linear model appears amenable to a variety of iterative solution methods. However, a significant challenge that arises is that in general the elements of $\boldsymbol{A}$ can be quite complicated in the form above, yet $\boldsymbol{A}$ is too large to store for typical problem sizes. Most iterative algorithms require matrix-vector multiplication by $\boldsymbol{A}$ and its transpose; there are fast algorithms for these operations (without storing $\boldsymbol{A}$ explicitly) in many special cases of interest [10], [11].

Thus far we have allowed the phase function $\phi(\vec{r}_j, t_i)$ to be quite general, without the traditional focus on "Fourier encoding." Recently there has been interest in investigating nonlinear magnetic field variations $B_z(\vec{r}, t)$ in (4), and reconstruction algorithms have been proposed that use much of the generality in (16) [12], [13]. These are currently specialized research topics, so we now focus on the more common case of linear field gradients.

FOURIER ENCODING

In typical MR scanners, the longitudinal component of the applied field $B_z(\vec{r}, t)$ in (4) consists of three components

$$B_z(\vec{r}, t) = B_0 + \Delta B_0(\vec{r}) + \vec{G}(t) \cdot \vec{r}. \tag{19}$$

The constant B_0 denotes the advertised field strength of the main static field. The function $\Delta B_0(\vec{r})$ denotes the spatial deviations of the field strength from this nominal value. This function is often called a field map, and in general, it is unknown, but it can be estimated by suitable types of acquisitions and data processing methods [14]. The field gradients $\vec{G}(t) = \vec{G}_x(t)\vec{j} + \vec{G}_y(t)\vec{j} + \vec{G}_z(t)\vec{j}$ consist of three user-controlled functions that are the historical key to providing spatial information in standard MR imaging. Many different types of MR scans are possible by changing $\vec{G}(t)$.

Substituting (19) into (11) using $\omega_0 \triangleq \gamma B_0$ and simplifying yields

$$\phi(\vec{r}, t) = \int_0^t \gamma\, \Delta B_0(\vec{r}) + \gamma \vec{G}(t) \cdot \vec{r}\, \mathrm{d}t$$

or equivalently

$$\mathrm{e}^{-i\phi(\vec{r},t)} = \mathrm{e}^{-i\Delta\omega_0(\vec{r})t}\, \mathrm{e}^{-i2\pi\vec{k}(t)\cdot\vec{r}}, \qquad (20)$$

where $\Delta\omega_0(\vec{r}) \triangleq \gamma\, \Delta B_0(\vec{\gamma})$ denotes the off-resonance frequency and the k-space trajectory is defined by

$$\vec{k}(t) \triangleq \frac{1}{2\pi}\int_0^t \gamma \vec{G}(t)\, \mathrm{d}t. \qquad (21)$$

Usually the phase accrual $\mathrm{e}^{-i\Delta\omega_0(\vec{r})t}$ due to off resonance is undesirable and can distort reconstructed images if ignored. Therefore some image reconstruction methods, particularly in fMRI, account for its effects [10]. In some cases, the map $\Delta\omega_0(\vec{r})$ is found from a separate "prescan," in other cases it is estimated jointly with f [15]. In chemical shift imaging, e.g., to separate fat and water components, the term $\Delta\omega_0(\vec{r})$ includes both useful information about the chemical shift effect as well as the undesirable variations due to field inhomogeneity [16].

For the linear field gradients (19), substituting (20) into (17) yields simpler expressions for the system matrix

$$a_{lij} \approx c_l(\vec{r}_j)\mathrm{e}^{-z(\vec{r}_j)t_i}\mathrm{e}^{-i2\pi\vec{k}(t_i)\cdot\vec{r}_j}, \qquad (22)$$

where we define the "rate map" $z(\vec{r})$ by combining the relaxation and field maps

$$z(\vec{r}) \triangleq 1/T_2^*(\vec{r}) + i\Delta\omega_0(\vec{r}). \qquad (23)$$

When this rate map is assumed to be zero, i.e., if relaxation and off resonance are ignored, then a_{lij} is the product of a Fourier encoding matrix having elements $\mathrm{e}^{-i2\pi\vec{k}(t_i)\cdot\vec{r}_j}$ with a diagonal sensitivity encoding matrix having elements $c_l(\vec{r}_j)$.

If the k-space sample locations $\vec{k}(t_i)$ lie on an appropriate subset of a Cartesian grid, then FFT operations provide efficient multiplication by A and its transpose. If non-Cartesian k-space sampling is used, then a nonuniform FFT (NUFFT) is needed [17].

When $z(\vec{r})$ in (22) is nonzero, then the elements (22) no longer correspond to a standard Fourier transform. Approximations are needed to provide fast computation of matrix-vector products. In particular, often one can approximate the exponentials in (17) using an additively separable form

$$\mathrm{e}^{-z(\vec{r}_j)t_i} \approx \sum_k b_{ik} c_{kj}$$

for various choices for the basis functions b_{ik} and coefficients c_{kj} [11]. With this type of approximation, we can rewrite matrix-vector multiplication as follows:

$$[A_l f]_i \approx \sum_k b_{ik} \sum_{j=1}^{N} (c_{kj}\, c_l(\vec{r}_j) f_j)\mathrm{e}^{-i2\pi\vec{k}(t_i)\cdot\vec{r}_j}.$$

The inner sum is simply a FFT or NUFFT so this approach is relatively fast. Free software for this is available [18].

RECONSTRUCTION COST FUNCTION

Having specified the linear model (18), we now turn to solution methods. Because the noise in MRI measurements is Gaussian, a natural approach is to estimate f by minimizing a regularized least-squares cost function

$$\hat{f} = \arg\min_f \Psi(f),\ \ \Psi(f) \triangleq \ \|y - Af\|^2 + \beta \mathrm{R}(f). \qquad (24)$$

For a single coil, the noise variance in the k-space data is white (uncorrelated with uniform variance), so the usual Euclidian norm $\|\cdot\|$ is appropriate. For parallel MRI, noise is stationary across time samples (i), but the norm should include the inverse of the $L \times L$ covariance matrix Σ in (13) that describes the noise correlation between receive coils [7].

If the k-space samples lie on an equally spaced grid (Cartesian sampling) with appropriate sample spacings relative to the object field of view, and if the rate map $z(\vec{r})$ is zero (i.e., we ignore relaxation and field inhomogeneity), and if we consider just a single coil $(L = 1)$ and treat the sensitivity pattern as uniform, i.e., $c_1(\vec{r}) = 1$, then the system matrix A_l is orthogonal. In this special case, no regularization is needed and $A^{-1} = 1/NA'$ and the solution is simply $\hat{f} = 1/NA'y$, which can be evaluated by an inverse FFT. This is the most common MR image-reconstruction method. However, if any of these conditions do not hold, then typically the system matrix A is not well conditioned, and the unregularized LS solution can lead to undesirable noise amplification. To avoid this problem, some form of regularization is needed.

REGULARIZATION

An open problem in most image reconstruction problems, including MRI, is how to best choose the regularizer $\mathrm{R}(f)$. If this term is not included, then the image estimate $\hat{f}$ will suffer from noise and artifacts for under-sampled and/or non-Cartesian data, because this inverse problem is ill conditioned. The approach for iterative reconstruction that has been adopted in commercial positron emission tomography scanners is to use an unregularized algorithm, initialize it with a uniform image, stop iterating just as the image gets unacceptably noisy, and then perhaps apply a bit of post-filtering to reduce the noise. One could adopt a similar approach for MR imaging. However, introducing regularization can ensure that the iterative algorithm converges to a stable image and can enforce prior information that improves image quality particularly for under-sampled data.

The simplest choice is Tikhonov regularization $\mathrm{R}(f) = \|f\|^2$ or $\mathrm{R}(f) = \|f - \bar{f}\|^2$, where $\bar{f}$ is some prior or reference image (possibly zero). The disadvantage of this choice is that it biases the estimate towards the reference image $\bar{f}$. In particular, if the reference image is zero, then all pixel values in $\hat{f}$ are diminished towards zero, possibly reducing contrast.

Another choice is a quadratic roughness penalty function, which in one-dimensional (1-D) would be written

$$\mathrm{R}(f) = \sum_{j=2}^{N} |f_j - f_{j-1}|^2. \tag{25}$$

This choice biases the reconstruction towards a smooth image where neighboring pixel values are similar. It is convenient for minimization [10], but it has the drawback of smoothing image edges, particularly if the regularization parameter β in (24) is too large. One can prove that using (25) guarantees that the cost function (24) has a unique minimizer.

RECENTLY IT HAS BECOME VERY POPULAR TO EXPRESS PRIOR INFORMATION IN TERMS OF SOME TYPE OF SPARSITY OF THE OBJECT.

More recently, total variation methods have been investigated for MR image reconstruction [19]. In 1-D, these methods replace the squared differences between neighboring pixels above with absolute differences

$$\mathrm{R}(f) = \sum_{j=2}^{N} |f_j - f_{j-1}|. \tag{26}$$

In 2-D continuous space, the analogous functional is

$$\int \|\nabla f\| \mathrm{d}\vec{r} = \iint \sqrt{\left|\frac{\partial}{\partial x} f(\vec{r})\right|^2 + \left|\frac{\partial}{\partial y} f(\vec{r})\right|^2} \, \mathrm{d}x \, \mathrm{d}y.$$

The advantage of this type of regularization is that it biases the reconstructed image towards a piecewise smooth image, instead of a globally smooth image, thereby better preserving image edges. However it is harder to minimize and can lead to the appearance of "blocky" texture in images. Numerous alternatives of the form

$$\mathrm{R}(f) = \sum_{j=2}^{N} \psi(f_j - f_{j-1})$$

for various choices of the "potential function" $\psi(\cdot)$ have been proposed in the imaging literature. Many of these compromise between the quadratic case (25) and the absolute difference case (26), for example the hyperbola

$$\psi(t) = \sqrt{1 + |t/\delta|^2} - 1 \tag{27}$$

is approximately quadratic near zero, which aids noise reduction, yet approximately linear away from zero, which helps preserve edges.

ALGORITHMS

Iterative algorithms are needed to minimize (24). For differentiable regularizers such as (25), the conjugate gradient algorithm is a natural choice [10]. For nondifferentiable regularizers like (26), more sophisticated algorithms are needed and this is an active research area [20].

RECONSTRUCTION CHALLENGES

Although a variety of useful problems can be solved in MRI using the formulation (24), there are numerous challenges that provide research opportunities.

REGULARIZATION PARAMETER SELECTION

A practical challenge with regularized methods is selection of the regularization parameter β in (24). For quadratic regularization, there is a well-developed theory for choosing β in terms of the desired spatial resolution properties of the reconstructed image [21]. This theory extends readily to MR imaging with reasonably well sampled trajectories (and to parallel imaging with reasonable acceleration factors) for which the point spread function (PSF) of the reconstructed image is relatively close to a Kronecker impulse so that simple measures like full width at half maximum (FWHM) are reasonable resolution metrics. For highly undersampled trajectories, the PSF can have "heavy tails" due to aliasing effects, and more investigation is needed to extend the above methods to MR applications.

For nonquadratic regularization such as the total variation method (26), the analysis in [21] is inapplicable so one must resort to other methods for choosing β. Statisticians often use cross validation for choosing regularization parameters, with a goal of finding the parameter that minimizes the mean-squared error (MSE) between $\hat{f}$ and the unknown f. However, MSE is the sum of variance and bias squared, and where bias is related to spatial resolution and artifacts, and it is unclear whether an equal weighting of noise variance and bias (squared) is optimal from an image-quality perspective in medical imaging.

Another method for choosing β is the "L-curve" method. This method is expensive because it requires evaluating $\hat{f}$ for several values of β, and it has some theoretical deficiencies [22].

In summary, choosing β for nonquadratic regularization remains a nontrivial issue in most ill-posed imaging problems including MRI, and remains an active research area [23].

PARTIAL K-SPACE METHODS

If the object $f(\vec{r})$ were real, then its Fourier transform would be Hermitian symmetric so in principle only half of k-space would need to be sampled. In practice, the magnetization (3) is complex due to a variety of physical effects. However, in many cases the phase of $\mathrm{M}(\vec{r}, t)$ can be assumed to be a smooth function. This property has led to a variety of partial k-space methods where one samples a bit more than half of k-space, then estimates the phase from the central portion of k-space (corresponding to low spatial frequencies), and then uses this estimated phase to reconstruct the entire image [24]. Such methods are used routinely in many types of MR scans.

UNDER-SAMPLED K-SPACE DATA

The need for some type of regularization is essential when the k-space data is under sampled, i.e., when the number of measurements Ln_d is less than the number of unknown voxels N. In MRI, the scan time is roughly proportional to the number of measurements, so collecting fewer samples can reduce scan time, which is particularly desirable in dynamic imaging.

In the broader field of tomographic image reconstruction, there is a long history of using prior information, such as assuming objects are piecewise smooth, to reconstruct images from an under-sampled set of projection views, e.g., [25]. Many of these methods involve cost functions of the form (24) with a suitable system matrix A for the application and appropriate regularizers $\mathrm{R}(f)$ that capture prior information about the object.

THE LINEAR IMAGE RECONSTRUCTION PROBLEM IS JUST ONE OF MANY ESTIMATION PROBLEMS OF INTEREST IN MRI.

Recently it has become very popular to express prior information in terms of some type of sparsity of the object. Notions of sparsity have deep roots in statistical signal processing [26]. Sparsity is especially apparent in MR angiography. The moniker of compressed sensing or compressive sampling has become widespread for such techniques, and recently entire sessions at MR conferences have been devoted to this topic [20]. Some compressed sensing formulations ignore the noise in the data. In the presence of noise, a typical formulation is

$$\arg\min_f \| \Psi f \|_1 \ \ \text{s.t.} \ \ \| y - Af \|_2 \le \epsilon,$$

where Ψ transforms the image f into a domain (such as wavelet coefficients) where one postulates that the signal is sparse.

Often this optimization problem is solved using a Lagrange multiplier approach

$$\arg\min_f \| y - Af \|_2^2 + \beta \| \Psi f \|_1,$$

which corresponds to a particular regularizer in (24). Rarely is the ℓ_1 norm implemented exactly; in practice usually a continuously differentiable approximation is used, such as

$$\| v \|_1 \approx \sum_i \left(\sqrt{|v_i|^2 + \delta^2} - \delta \right) \qquad (28)$$

for some small value of $\delta > 0$. This approximation is equivalent to the hyperbola (27) used frequently for edge-preserving image reconstruction. Nonconvex methods that enforce sparsity even more strongly are also under investigation. In the usual case where A corresponds to an under-sampled discrete Fourier transform (DFT), a variety of algorithms are available that have numerous potential applications in MR [20]. Challenges with this approach include choosing the sparsifying transform Ψ and regularization parameters β and δ appropriately. Furthermore, when δ is small, the regularizer (28) has very high curvature near zero, which can slow convergence.

NONLINEAR RECONSTRUCTION PROBLEMS

The linear image reconstruction problem (24) is just one of many estimation problems of interest in MRI. Returning to the elements of the system matrix (22), there has been research on estimating essentially every component therein, as summarized below.

FIELD MAP ESTIMATION

For scans with long readout times, the effect of field inhomogeneity $\Delta\omega_0$ in (22) is important. In practice, the field map $\omega(\vec{r})$ is not known a priori but rather it must be estimated from noisy MR scans. One can examine the phase differences between two scans having different echo times to determine $\Delta\omega_0$. If these two scans have short readouts, then there are simple image-domain methods for estimating $\Delta\omega_0$, which is known as B_0 field mapping [14]. Errors in the field map estimates may cause artifacts in reconstructed images that are based on models like (22).

In addition, object motion that occurs between the field map scans and subsequent scans of interest, e.g., in fMRI, will lead to an inconsistency between the actual scan data and the assumed model (22) used by the reconstruction algorithm. This possibility has motivated the development of dynamic field mapping methods that estimate the field map separately for each frame in a dynamic study, e.g., [15]. For scans with long readout durations, the appearance of $\Delta\omega_0$ in a complex exponential in (22) makes this a somewhat complicated nonlinear estimation problem.

RELAXATION MAP ESTIMATION

In some MR applications, it is useful to estimate tissue relaxation parameters, particularly T_2 or T_2^*, on a pixel-by-pixel basis. One approach to measuring such relaxation parameters is to acquire a "baseline" scan of the object and then acquire one or more additional scans having different echo times. One then reconstructs images from each of those scans and then performs linear regression on a voxel-by-voxel basis using the logarithm of the image voxel values. This approach can be adequate if the readout durations are sufficiently small. But for acquisitions with long readouts, the effect of time t_i in the $\mathrm{e}^{-z(\vec{r}_j)t_i}$ in (22) should be considered, i.e., we should account for relaxation during the signal readout. This requires methods that estimate the relaxation map directly from the k-space data. These are more challenging nonlinear estimation problems because T_2^* appears in an exponent in (22). Several methods for jointly estimating T_2^*, $\Delta\omega_0$, and $f(\vec{r})$ have been investigated [27].

SENSITIVITY MAP ESTIMATION

The coil sensitivity patterns $c_l(\vec{r})$ in (22) also must be determined for parallel imaging based on sensitivity encoding. Normally this is done by acquiring well-sampled data both with local receive coils and with a reference body coil and dividing the two [7]. Acquiring the extra reference data can be inconvenient, so normalizing by the square root of the sum of squares of the local receive coils is also used. A variety of other estimation methods have been proposed, including methods that jointly estimate the sensitivity maps $\{c_l(\vec{r})\}$ and the image $f(\vec{r})$ [28]. Note that if $f(\vec{r})$ were known, then the problem of estimating $c_l(\vec{r})$ would be a linear estimation problem because $c_l(\vec{r})$ appears as a linear scaling in (22). But when both $f(\vec{r})$ and $c_l(\vec{r})$ are to be estimated, the model is bilinear because $f(\vec{r})$ and $c_l(\vec{r})$ appear as a product in (10). This complicates joint estimation.

DESPITE OVER THREE DECADES OF MR RESEARCH, THERE REMAIN CHALLENGING AND INTRIGUING PROBLEMS IN MR IMAGE RECONSTRUCTION.

TRAJECTORY MAPPING

The k-space trajectory $\vec{k}(t_i)$, defined as an integral of the gradient waveforms in (21), should be calibrated carefully to ensure that the system model (22) is accurate. In practice, the field gradients induced by the gradient coils in the scanner are not exactly proportional to the waveforms applied to those coils due to eddy currents. Therefore the physical k-space trajectory realized in the system can depart somewhat from the desired k-space trajectory. These differences can degrade the reconstructed image, particularly for non-Cartesian trajectories with long readout durations. Therefore, a variety of techniques have been developed for mapping the actual k-space trajectory experimentally.

WITHIN-VOXEL GRADIENTS

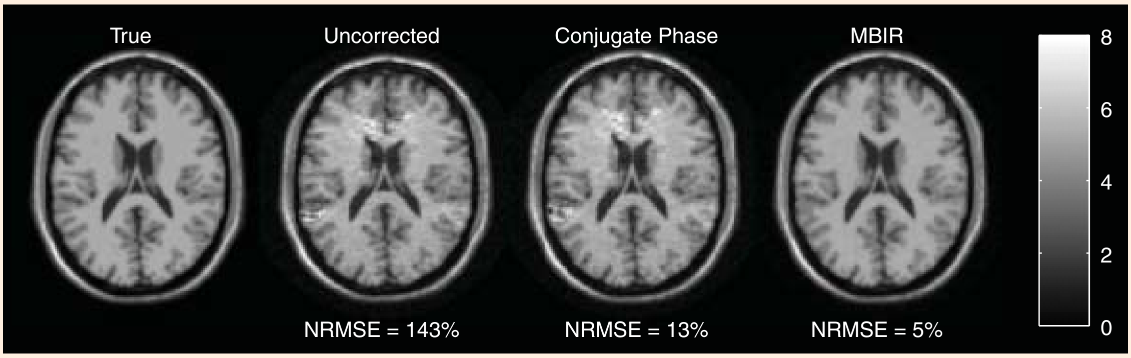
The model (23) treats the field inhomogeneity within each voxel as being a constant, ignoring within-voxel gradients of the off-resonance map. However, these gradients can be significant in functional magnetic resonance imaging (fMRI) based on the BOLD effect [29]. Accurate reconstruction of signals near air-tissue interfaces requires compensation for these within-voxel gradients, which complicates the reconstruction method [30].

EXAMPLE

To illustrate the capabilities of model-based image reconstruction methods for MRI, we simulated k-space data for a four-shot EPI sequence with matrix size 128 × 128 and 5 μs sampling so the readout duration was 27.3 μs per shot. The field map $\Delta B_0(\vec{r})$ appears in Figure 2 of [14] and is based on a brain slice above the sinuses and ear canals where susceptibility effects occur. Figure 1 shows the true image used in the simulations and images from three different reconstruction methods. The "uncorrected" reconstruction simply uses an inverse 2-D FFT, with no consideration of field inhomogeneity. The field inhomogeneity causes spatial distortion in the read-out (vertical) direction (that increases NRMSE dramatically), as well as significant intensity artifacts above the ears and sinuses where the susceptibility effects are largest. The classical conjugate phase reconstruction method, which corresponds to $A'y$ in this single-coil case, reduces the spatial distortion but the intensity artifacts persist. Applying 15 iterations of a conjugate gradient algorithm with a monotonic line search [11] to the cost function (24) with the edge-preserving hyperbola (27) yields the right-most image in Figure 1. This model-based image reconstruction method yields the lowest RMS error, but it requires about 30 times more computation than the noniterative conjugate phase method [11] because each iteration requires multiplication by A and A'. The software that generated this figure is available online [18].

SUMMARY

Image reconstruction is not a single problem in MRI but rather is a wide family of problems depending on what physical



[FIG1] Comparison of model-based image reconstruction with convention methods.

effects are included in the signal model. The most widely studied case, particularly in the signal processing community, is when nearly all physical effects are disregarded and the system model consists solely of sampled of the Fourier transform of the object. This basic model is amenable to familiar signal processing tools and is applicable to many MR scans. But there are also many interesting applications where other physical effects are relevant, and model-based methods that account for those effects are proving to be beneficial for improving image quality. Model-based methods themselves depend on estimates of a variety of model parameters, leading to interesting problems where those parameters are determined either by separate calibration scans or by jointly estimating the image and those parameters. Despite over three decades of MR research, there remain challenging and intriguing problems in MR image reconstruction.

ACKNOWLEDGMENT

This work was supported in part by NIH grants CA87634 and NS058576.

AUTHOR

Jeffrey A. Fessler (fessler@umich.edu) received the B.S.E.E. degree from Purdue University in 1985, the M.S.E.E. degree from Stanford University in 1986, and the M.S. degree in statistics from Stanford University in 1989. From 1985 to 1988, he was a National Science Foundation Graduate Fellow at Stanford, where he earned a Ph.D. degree in electrical engineering in 1990. He has worked at the University of Michigan since thenwhere he is now a professor in the Departments of Electrical Engineering and Computer Science, Radiology, and Biomedical Engineering. He is a Fellow of the IEEE. He received the Francois Erbsmann Award. He is an associate editor for *IEEE Transactions on Medical Imaging* and was an associate editor for *IEEE Transactions on Image Processing* and *IEEE Signal Processing Letters*. He was cochair of the 1997 SPIE Conference on Image Reconstruction and Restoration, technical program cochair of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI), and general chair of ISBI 2007. His research interests are in statistical aspects of imaging problems.

REFERENCES

[1] M. Bydder, A. A. Samsonov, and J. Du, "Evaluation of optimal density weighting for regridding," *Magn. Reson. Imag.,* vol. 25, no. 5, pp. 695–702, June 2007.

[2] G. A. Wright, "Magnetic resonance imaging," *IEEE Signal Processing Mag.,* vol. 14, no. 1, pp. 56–66, Jan. 1997.

[3] Z.-P. Liang and P. C. Lauterber, *Principles of Magnetic Resonance Imaging.* New York: IEEE Press, 2000.

[4] W. A. Grissom, D. Xu, A. B. Kerr, J. A. Fessler, and D. C. Noll, "Fast large-tip-angle multidimensional and parallel RF pulse design in MRI," *IEEE Trans. Med. Imag.,* vol. 28, no. 10, pp. 1548–1559, Oct. 2009.

[5] P. Mansfield and I. L. Pykett, "Biological and medical imaging by NMR," *J. Magn. Reson., vol.* 29, no. 2, pp. 355–373, Feb. 1978.

[6] P. B. Roemer, W. A. Edelstein, C. E. Hayes, S. P. Souza, and O. M. Mueller, "The NMR phased array," *Magn. Reson. Med.,* vol. 16, no. 2, pp. 192–225, Nov. 1990.

[7] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boe-siger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.,* vol. 42, no. 5, pp. 952–962, Nov. 1999.

[8] A. Macovski, "Noise in MRI," *Magn. Reson. Med.,* vol. 36, no. 3, pp. 494–497, Sept. 1996.

[9] Y. Censor, "Finite series expansion reconstruction methods," *Proc. IEEE,* vol. 71, no. 3, pp. 409–419, Mar. 1983.

[10] B. P. Sutton, D. C. Noll, and J. A. Fessler, "Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities," *IEEE Trans. Med. Imag.,* vol. 22, no. 2, pp. 178–188, Feb. 2003.

[11] J. A. Fessler, S. Lee, V. T. Olafsson, H. R. Shi, and D. C. Noll, "Toeplitz-based iterative image reconstruction for MRI with correction for magnetic field inhomogeneity," *IEEE Trans. Signal Processing,* vol. 53, no. 9, pp. 3393–3402, Sept. 2005.

[12] J. Hennig, A. M. Welz, G. Schultz, J. Korvink, Z. Liu, O. Speck, and M. Zaitsev, "Parallel imaging in non-bijective, curvilinear magnetic field gradients: A concept study," *Magn. Reson. Mater. Biol. Phys. Med.,* vol. 21, no. 1–2, pp. 5–14, Mar. 2008.

[13] J. Stockmann, P. Ciris, and R. T. Constable, "Efficient "O-space" parallel imaging with higher-order encoding gradients and no phase encoding," in *Proc. Int. Soc. Magnetic Resonance in Medicine,* 2009, p. 761.

[14] A. K. Funai, J. A. Fessler, D. T. B. Yeo, V. T. Olafsson, and D. C. Noll, "Regularized field map estimation in MRI," *IEEE Trans. Med. Imag.,* vol. 27, no. 10, pp. 1484–1494, Oct. 2008.

[15] B. P. Sutton, D. C. Noll, and J. A. Fessler, "Dynamic field map estimation using a spiral-in/spiral-out acquisition," *Magn. Reson. Med.,* vol. 51, no. 6, pp. 1194–1204, June 2004.

[16] W. Dixon, "Simple proton spectroscopic imaging," *Radiology,* vol. 153, no. 1, pp. 189–194, Oct. 1984.

[17] J. A. Fessler and B. P. Sutton, "Nonuniform fast Fourier transforms using min-max interpolation," *IEEE Trans. Signal Processing,* vol. 51, no. 2, pp. 560–574, Feb. 2003.

[18] J. A. Fessler. MATLAB tomography toolbox, 2004 [Online]. Available: http://www.eecs.umich.edu/~fessler

[19] K. T. Block, M. Uecker, and J. Frahm, "Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint," *Magn. Reson. Med.,* vol. 57, no. 6, pp. 1086–1098, June 2007.

[20] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Mag.,* vol. 25, no. 2, pp. 72–82, Mar. 2008.

[21] J. A. Fessler and W. L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs," *IEEE Trans. Image Processing,* vol. 5, no. 9, pp. 1346–1358, Sept. 1996.

[22] C. R. Vogel, "Non-convergence of the L-curve regularization parameter selection method," *Inverse Probl.,* vol. 12, no. 4, pp. 535–547, Aug. 1996.

[23] S. Ahn and R. M. Leahy, "Analysis of resolution and noise properties of nonquadratically regularized image reconstruction methods for PET," *IEEE Trans. Med. Imag.,* vol. 27, no. 3, pp. 413–424, Mar. 2008.

[24] D. C. Noll, D. G. Nishimura, and A. Macovski, "Homodyne detection in magnetic resonance imaging," *IEEE Trans. Med. Imag.,* vol. 10, no. 2, pp. 154–163, June 1991.

[25] A. H. Delaney and Y. Bresler, "Globally convergent edge-preserving regularized reconstruction: An application to limited-angle tomography," *IEEE Trans. Image Processing,* vol. 7, no. 2, pp. 204–221, Feb. 1998.

[26] G. Harikumar and Y. Bresler, "A new algorithm for computing sparse solutions to linear inverse problems," in *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing,* 1996, vol. 3, pp. 1331–1334.

[27] V. T. Olafsson, D. C. Noll, and J. A. Fessler, "Fast joint reconstruction of dynamic R_2^* and field maps in functional MRI," *IEEE Trans. Med. Imag.,* vol. 27, no. 9, pp. 1177–1188, Sept. 2008.

[28] L. Ying and J. Sheng, "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)," *Magn. Reson. Med.,* vol. 57, no. 6, pp. 1196–1202, June 2007.

[29] D. C. Noll, J. A. Fessler, and B. P. Sutton, "Conjugate phase MRI reconstruction with spatially variant sample density correction," *IEEE Trans. Med. Imag.,* vol. 24, no. 3, pp. 325–336, Mar. 2005.

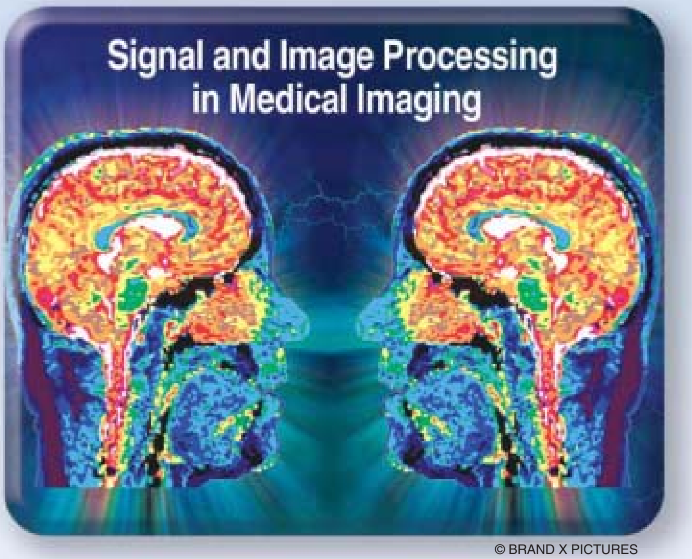
[30] J. A. Fessler and D. C. Noll, "Model-based MR image reconstruction with compensation for through-plane field inhomogeneity," in *Proc. IEEE Int. Symp. Biomedical Imaging,* 2007, pp. 920–923.

[SP]

[Leslie Ying and Zhi-Pei Liang]

Parallel MRI Using Phased Array Coils

[Multichannel sampling theory meets spin physics]



© BRAND X PICTURES

Magnetic resonance imaging (MRI) is a relatively slow imaging technique that has limited its application to imaging of time-varying objects. Developing fast MRI methods has been an active research area for the last three decades. Recently, parallel imaging using phased array coils has provided another avenue to significantly speed up the MRI process. In this article, we describe parallel MRI from a signal processing perspective, invoking the multichannel sampling theory (and filter bank theory). We review several basic reconstruction algorithms and discuss some practical issues and outstanding signal processing problems.

INTRODUCTION

MRI is a tomographic imaging technique based on the well-known nuclear magnetic resonance (NMR) phenomenon. An intuitive understanding of the NMR phenomenon can be gained by considering the ^1H nuclei (protons) in an object to be imaged. The proton is a positively charged particle that has an intrinsic angular momentum (called spin) and a microscopic magnetic field surrounding it (characterized by a magnetic moment vector). Under the thermal equilibrium condition, the object displays no macroscopic magnetism due to destructive interference among all the magnetic moment vectors. However, when placed in an external magnetic field ($\vec{B}_0$), the protons will be polarized, giving rise to a bulk magnetization vector $\vec{M}$ pointing in the direction of $\vec{B}_0$. The thermal equilibrium value of $\vec{M}$ is given, to a first-order approximation by

$$M_0 = \frac{\gamma^2 \hbar^2 B_0 N_s}{4 k_B T_s}, \tag{1}$$

where γ is the gyromagnetic ratio of proton (2.675×10^8 rad/s/T), $\hbar$ is the Planck constant h (6.6×10^{-34} J-s) divided by 2π, k_B is the Boltzmann constant (1.38×10^{-23} J/K), N_s is the total number of polarized protons in the object, and T_s is the absolute temperature of the object. When the "magnetized" object is further excited by another field oscillating at the Larmor frequency ($\omega_0 = \gamma B_0$) of the protons, the bulk magnetization $\vec{M}$ will be tilted away from the $\vec{B}_0$ field. The tilted $\vec{M}$ will then precess about the $\vec{B}_0$ field (known as free precession), and consequently induce a voltage signal (known as the NMR signal) in a radio frequency (RF) receiver coil according to Faraday's law of induction

Digital Object Identifier 10.1109/MSP.2010.936731

1053-5888/10/$26.00©2010IEEE

$$v(t) = -\frac{d}{dt}\int \vec{M}(\vec{r}, t) \cdot \vec{B}_c(\vec{r}) d\vec{r}, \qquad (2)$$

where $\vec{B}_c(\vec{r})$ represents the detection sensitivity of the receiver coil at spatial location $\vec{r}$.

In conventional MRI systems, both signal excitation and detection are performed using a single RF channel, which is assumed to have uniform sensitivity over the region of interest. In this case, the received NMR signal (after proper simplifications and processing such as filtering and demodulation) can be expressed as

$$s(t) = \int_{\mathrm{FOV}} \rho(\vec{r})\, e^{-i\Delta\omega t} d\vec{r}, \qquad (3)$$

where $\rho(\vec{r})$ is the spin (proton) density function such that $M_0 \propto \int \rho(\vec{r})\, d\vec{r}$, $\Delta\omega$ is the frequency of $\vec{M}$ precessing about the $\vec{B}_0$ field (in the rotating frame), and FOV denotes the field of view. Note that in (3), all the modulating factors (such as T_1-weighting, T_2-weighting, and diffusion-weighting, which are important for image contrast) are ignored for simplicity such that $\rho(\vec{r})$ can be taken as the desired image function. The time signal $s(t)$ can be mapped to the Fourier domain (commonly called k-space) if the signal is acquired in the presence of a gradient field (characterized by gradient vector $\vec{G} = (G_x, G_y, G_z)$ such that $\Delta\omega(\vec{r}) = \gamma\vec{G} \cdot \vec{r}$. In this case, (3) can be rewritten as

$$s(\vec{k}) = \int_{\mathrm{FOV}} \rho(\vec{r})\, e^{-i2\pi\vec{k}\cdot\vec{r}} d\vec{r}, \qquad (4)$$

where $\vec{k} = \gamma\vec{G}t/2\pi$. This equation is known as the Fourier imaging equation for MRI.

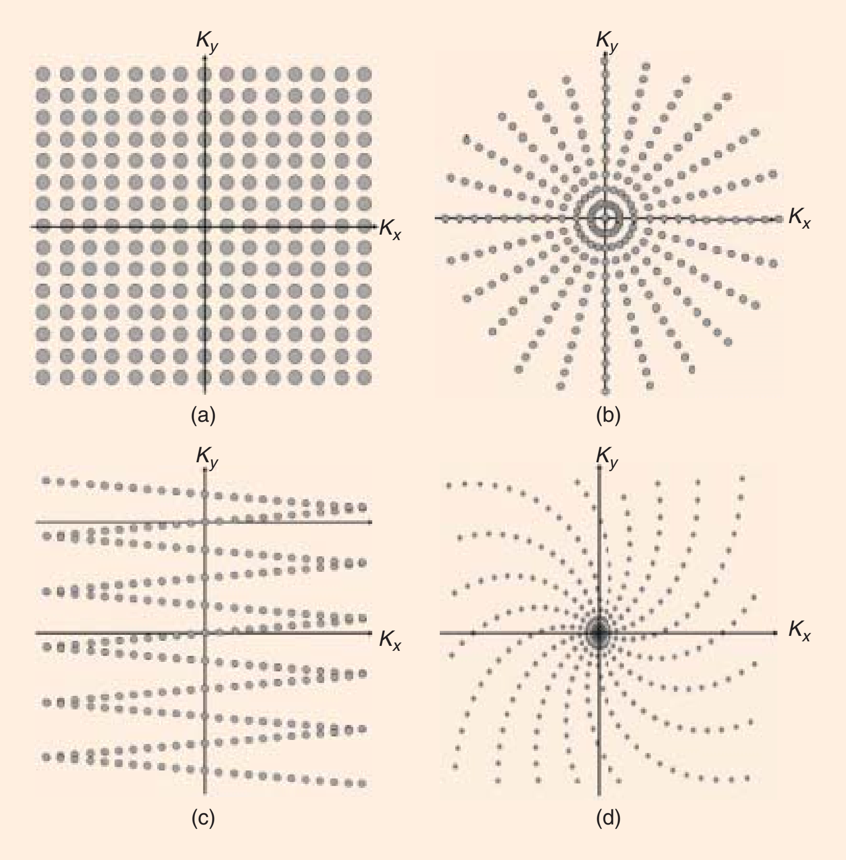
MR physics provides a lot of flexibility in sampling k-space. Some typical examples of k-space sampling are shown in Figure 1. In both Cartesian and polar sampling [Figure 1(a) and (b)], each line of data is commonly generated by one RF excitation and acquired in the presence of constant gradients, which often leads to long data acquisition time. The non-Cartesian trajectories in Figure 1(c) and (d) are generated using time-varying gradients and the data can be acquired more efficiently. Regardless of the sampling trajectories, sampling of k-space in conventional Fourier imaging is governed by Shannon's sampling theorem and the number of measurements needed increases exponentially with the dimension of the image function. This presents a significant practical problem for fast imaging. Sparser sampling of k-space can reduce imaging time for MRI, and parallel imaging using multichannel phased array coils is an effective method for doing so. Many methods exist to achieve sub-Nyquist sampling of k-space, such as those based on compressive sampling theory [1] and the theory of partially separable functions [2]. A review of these methods is beyond the scope of this article.

We next review parallel MRI in the context of Papoulis' multichannel sampling theorem.

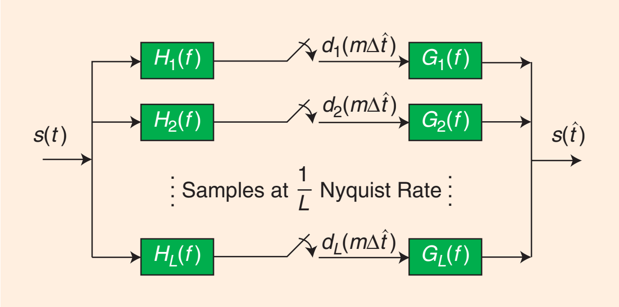
MULTICHANNEL SAMPLING THEORY

PAPOULIS' SAMPLING THEOREM

Consider a finite-energy signal $s(t)$ that is bandlimited to $|f| < B/2$. The Shannon sampling theorem [3] states that $s(t)$ is uniquely determined from its samples when the sampling interval is less than $\Delta t = 1/B$, where Δt is known as the Nyquist sampling interval and B the Nyquist rate. Papoulis introduced an important extension to Shannon's sampling theorem to address the multichannel sampling problem. The theorem states that $s(t)$ can be recovered exactly (under some conditions) from the samples of the output signals of L linear time-invariant (LTI) filters, sampled at $1/L$ of the Nyquist rate [4]. To state it more formally, let $s(t)$ be the common input to L LTI systems with frequency responses $H_1(f), H_2(f), \ldots, H_L(f)$, as shown in Figure 2. The outputs of these filters can be expressed as



[FIG1] Typical sampling patterns in k-space: (a) Cartesian, (b) polar, (c) zig-zag, and (d) spiral.



[FIG2] A diagram illustrating Papoulis' generalized sampling theorem.

$$s_\ell(t) = \int_{-B/2}^{B/2} S(f)\, H_\ell(f)\, e^{i2\pi ft} df, \qquad (5)$$

where $S(f)$ is the Fourier transform of $s(t)$. Papoulis showed [4] that $s(t)$ can be recovered from the samples of $s_\ell(t)$ taken at $1/L$ of the Nyquist rate of $s(t)$, denoted as $s_\ell(m\Delta\hat t)$. More specifically,

$$s(t) = \sum_{\ell=1}^{L} \sum_{m=-\infty}^{\infty} s_\ell(m\Delta\hat t)\, g_\ell(t - m\Delta\hat t), \qquad (6)$$

where $\Delta\hat t = L/B$, and $g_\ell(t)$ is an interpolation function obtained from

$$\sum_{\ell=1}^{L} H_\ell(f) \sum_{m=-\infty}^{\infty} g_\ell(t - m\Delta\hat t)\, e^{i2\pi fm\Delta\hat t} = e^{i2\pi ft}, \qquad (7)$$

for $f \in (-B/2, B/2)$. It can be shown [5]

$$g_\ell(t) = \int_{-B/2}^{B/2} G_\ell(f)\, e^{i2\pi ft} df, \qquad (8)$$

where the $G_\ell(f)$ for $f \in (-B/2, B/2)$ are the solutions of the following equation:

$$\mathbf{H}(f)\, \vec G(f) = \Delta\hat t\, \vec e, \qquad (9)$$

where

$$\mathbf{H}^T(f) = \begin{bmatrix} H_1(f) & H_1(f-\hat B) & \cdots & H_1(f-(L-1)\,\hat B) \\ H_2(f) & H_2(f-\hat B) & \cdots & H_2(f-(L-1)\,\hat B) \\ \vdots & \vdots & & \vdots \\ H_L(f) & H_L(f-\hat B) & \cdots & H_L(f-(L-1)\,\hat B) \end{bmatrix},$$

$$\vec G(f) = \begin{bmatrix} G_1(f) \\ G_2(f) \\ \vdots \\ G_L(f) \end{bmatrix}, \text{ and } \vec e = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

with $\hat B = B/L$. Equivalently, in the reduced frequency band $f \in (B/2 - \hat B, B/2)$, the $G_\ell(f)$ are given by

$$\mathbf{H}(f)\, \mathbf{G}(f) = \Delta\hat t\, \mathbf{I}, \qquad (10)$$

where $\mathbf{I}$ is an identity matrix, and $\mathbf{G}(f) = [\vec G(f), \vec G(f-\hat B), \cdots, \vec G(f-(L-1)\hat B)]$.

The signal $s(t)$ can also be reconstructed in the frequency domain by recovering its spectrum $S(f)$. Specifically, for $f \in (B/2 - \hat B, B/2)$, let

$$\vec S(f) = \begin{bmatrix} S(f) \\ S(f-\hat B) \\ \vdots \\ S(f-(L-1)\hat B) \end{bmatrix}, \text{ and } \vec S^a(f) = \begin{bmatrix} S_1^a(f) \\ S_2^a(f) \\ \vdots \\ S_L^a(f) \end{bmatrix},$$

where

$$\begin{aligned} S_\ell^a(f) &= \sum_{m=-\infty}^{\infty} s_\ell(m\Delta\hat t)\, e^{-i2\pi fm\Delta\hat t}, \\ &= \hat B \sum_{r=0}^{L-1} S(f - r\hat B)\, H_\ell(f - r\hat B). \end{aligned} \qquad (11)$$

We have from (11) that

$$\vec S^a(f) = \hat B \mathbf{H}^T(f) \vec S(f). \qquad (12)$$

The original signal spectrum can be recovered by

$$\vec S(f) = \Delta\hat t \mathbf{H}^{-T}(f) \vec S^a(f). \qquad (13)$$

Both the time-domain and frequency-domain reconstruction formulas indicate that $\mathbf{H}(f)$ must be full rank for $f \in (B/2 - \hat B, B/2)$ to assure unique perfect reconstruction for $s(t)$. In other words, perfect reconstruction of $s(t)$ requires that the L vectors $\{[H_\ell(f), H_\ell(f-\hat B), \ldots, H_\ell(f-(L-1)\,\hat B)]\}_{\ell=1}^{L}$ are linearly independent for each $f \in (B/2 - \hat B, B/2)$, which is a stronger condition than requiring $\{H_\ell(f)\}_{\ell=1}^{L}$ to be linearly independent functions.

CONNECTION TO DIGITAL FILTER BANKS

The multichannel sampling theorem can also be conveniently expressed in terms of digital filter banks [6] when we are only interested in the recovery of Nyquist samples (instead of the complete continuous function) as is the case with imaging. At the Nyquist sampling rate ($\Delta t = 1/B$), the continuous-time signal spectrum $S(f)$ for $|f| < B/2$ can be obtained from the discrete-time samples $s(n\Delta t)$ by

$$S(f) = \Delta t \sum_n s(n\Delta t)\, e^{-i2\pi fn\Delta t}. \qquad (14)$$

Substituting (14) into (5) yields

$$\begin{aligned} s_\ell(m\Delta\hat t) &= \Delta t \sum_n s(n\Delta t) \int_{-B/2}^{B/2} H_\ell(f)\, e^{i2\pi f(m\Delta\hat t - n\Delta t)} df, \\ &= \Delta t \sum_n s(n\Delta t)\, h_\ell(m\Delta\hat t - n\Delta t), \end{aligned}$$

$$= \Delta t (s * h_\ell)\,(m\Delta\hat{t}),$$

$$= \Delta t \downarrow_L (s * h_\ell)\,(n\Delta t), \qquad (15)$$

where * denotes convolution and $\downarrow_L$ denotes downsampling by a factor of L. Equation (15) indicates that $s_\ell(m\Delta\hat{t})$ (from the continuous time signal $s(t)$ filtered by analog filter $h_\ell(t)$ and sampled at $m\Delta\hat{t}$) are the same as the data obtained from the sequence $s(n\Delta t)$ filtered by a digital filter $\Delta t\, h_\ell(n\Delta t)$, followed by downsampling by a factor of L. This representation allows us to reconstruct the Nyquist samples $s(n\Delta t)$ using the digital filter bank theory, as illustrated in Figure 3, where the $\overline{H}_\ell(z)$ and $\overline{G}_\ell(z)$ are transfer functions (z transforms of filter coefficients) of the so-called analysis and synthesis filters, respectively. The reconstruction procedure can be expressed as [7]

$$s(n\Delta t) = \sum_{\ell=1}^{L}\sum_{m=-\infty}^{\infty} s_\ell(mL\Delta t)\, g_\ell((n-mL)\,\Delta t), \qquad (16)$$

where the $g_\ell(n\Delta t)$ are digital synthesis filter coefficients. The transfer function of the synthesis filters required for exact recovery of $s(n\Delta t)$ are given by

$$\overline{\mathrm{H}}(z)\,\vec{\overline{G}}(z) = L\vec{e}, \qquad (17)$$

where $\overline{\mathrm{H}}^T(z)$ is defined as

$$\begin{bmatrix} \overline{H}_1(z) & \overline{H}_1(zW^{-1}) & \cdots & \overline{H}_1(zW^{-(L-1)}) \\ \overline{H}_2(z) & \overline{H}_2(zW^{-1}) & \cdots & \overline{H}_2(zW^{-(L-1)}) \\ \vdots & \vdots & & \vdots \\ \overline{H}_L(z) & \overline{H}_L(zW^{-1}) & \cdots & \overline{H}_L(zW^{-(L-1)}) \end{bmatrix},$$

and

$$\vec{\overline{G}}(z) = \begin{bmatrix} \overline{G}_1(z) \\ \overline{G}_2(z) \\ \vdots \\ \overline{G}_L(z) \end{bmatrix},$$

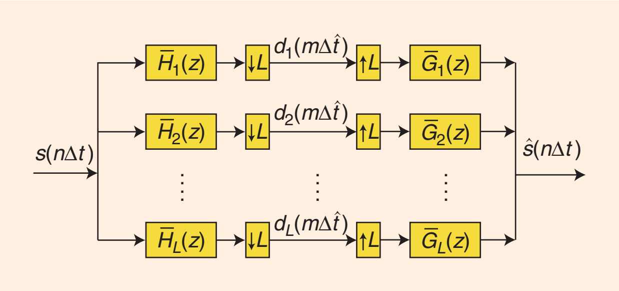
where $W = e^{i2\pi/L}$. It can be shown that the digital filters and their analog counterparts are related to each other by $\overline{\mathrm{H}}(e^{i2\pi f\Delta t}) = \mathrm{H}(f)$ and $\vec{\overline{G}}(e^{i2\pi f\Delta t}) = \vec{G}(f)/\Delta t$ for $f \in (-B/2, B/2)$.

MULTIDIMENSIONAL SAMPLING

Papoulis' sampling theorem can be extended to M dimensions [8]. Consider a finite-energy d-dimensional signal $s(\vec{t})$ for $\vec{t} = [t_1, t_2, \ldots, t_d]^T$, which is "bandlimited" to a d-dimensional cell $\mathcal{C}$. Passing the signal through L d-dimensional filters $H_1(\vec{f})$, $H_2(\vec{f})$, ..., and $H_L(\vec{f})$ for $\vec{f} = [f_1, f_2, \ldots, f_d]^T$ yields

$$s_\ell(\vec{t}) = \int_{\mathcal{C}} S(\vec{f})\, H_\ell(\vec{f})\, e^{-i2\pi\vec{f}\cdot\vec{t}}\, d\vec{f} \qquad (18)$$

for the ℓth channel. The multidimensional extension of Papoulis' sampling theorem enables perfect reconstruction of $s(\vec{t})$ from samples of $s_\ell(\vec{t})$ taken below the Nyquist rate. Specifically, let $\{\vec{v}_j\}_{j=1}^{d}$ define a sampling lattice in the d-dimensional $\vec{t}$-space (a two-dimensional (2-D) example is shown in Figure 4). $\mathrm{V}\vec{m}$ specifies a sampling of $\vec{t}$, where $\mathrm{V} = (\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d)$ and $\vec{m}$ is a d-dimensional vector of integers. It can be shown that if the sampling density of $\vec{t}$ is not lower than $1/L$ of the Nyquist density, we have (under some conditions on $H_\ell(\vec{f})$ similar to the one-dimensional (1-D) case)



[FIG3] A diagram of a digital filter bank.

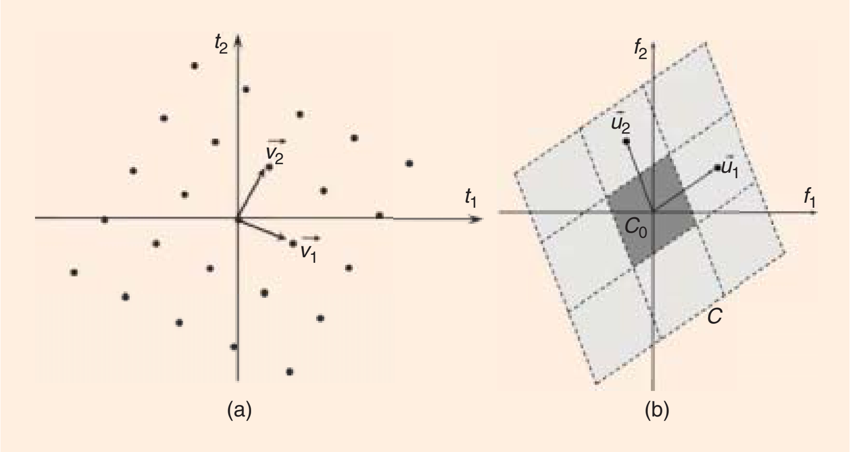
$$s(\vec{t}) = \sum_{\ell=1}^{L}\sum_{\vec{m}\in Z^d} s_\ell(\mathrm{V}\vec{m})\, g_\ell(\vec{t} - \mathrm{V}\vec{m}), \qquad (19)$$

where $g_\ell(\vec{t})$ is an interpolation function. The Fourier transform $G_\ell(\vec{f})$ of $g_\ell(\vec{t})$ is given by

$$\mathrm{H}(\vec{f})\vec{G}(\vec{f}) = |\det(\mathrm{V})|\,\vec{e}, \qquad (20)$$

where $\mathrm{G}(\vec{f}) = [G_1(\vec{f}), G_2(\vec{f}), \ldots, G_L(\vec{f})]^T$, and

$$\mathrm{H}^T(\vec{f}) = \begin{bmatrix} H_1(\vec{f}) & H_1(\vec{f} - \mathrm{U}\vec{q}_1) & \cdots & H_1(\vec{f} - \mathrm{U}\vec{q}_{L-1}) \\ H_2(\vec{f}) & H_2(\vec{f} - \mathrm{U}\vec{q}_1) & \cdots & H_2(\vec{f} - \mathrm{U}\vec{q}_{L-1}) \\ \vdots & \vdots & & \vdots \\ H_L(\vec{f}) & H_L(\vec{f} - \mathrm{U}\vec{q}_1) & \cdots & H_L(\vec{f} - \mathrm{U}\vec{q}_{L-1}) \end{bmatrix},$$



[FIG4] Two-dimensional sampling: (a) a sampling pattern and (b) subcells that are shifted replica of reference subcell $\mathcal{C}_0$ covering the spectrum $S(f_1, f_2)$ and folded together when $s(t_1, t_2)$ is sampled with the pattern in (a).

with $\mathbf{U} = \mathbf{V}^{-1}$. A detailed description of the multidimensional, multichannel sampling theorem can be found in [8].

WELL POSEDNESS

Assume that each undersampled measurement $s_\ell(m\Delta\hat{t})$ is corrupted by additive white noise with zero mean and variance σ_s^2 and there is no correlation between channels (correlated noise can be prewhitened). According to the multichannel sampling theory, the noise level of the interpolated data is [9]

$$\sigma^2(t) = \sigma_s^2 \sum_{\ell=1}^{L} \sum_{m=-\infty}^{\infty} |g_\ell(t - m\Delta\hat{t})|^2. \qquad (21)$$

In the frequency domain, we have

$$\sigma^2(f) = \frac{\sigma_s^2}{\Delta\hat{t}} \sum_{\ell=1}^{L} |G_\ell(f)|^2. \qquad (22)$$

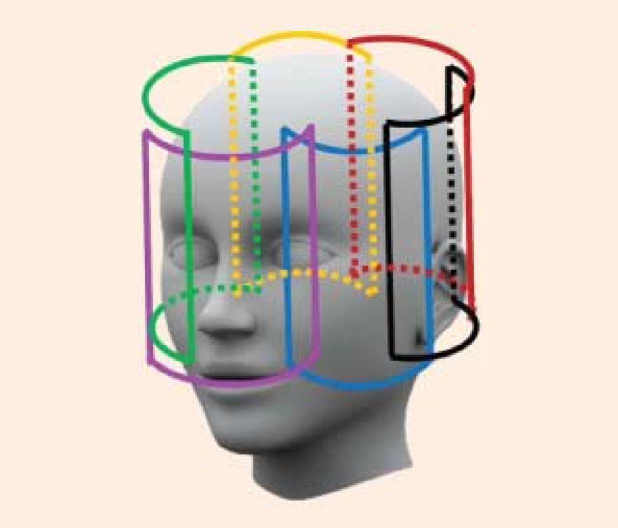
In considering the well posedness of the multichannel sampling problem, often the average noise level is evaluated, which is given by

$$\overline{\sigma}^2 = \frac{\sigma_s^2}{\Delta\hat{t}} \sum_{\ell=1}^{L} \int_{-\infty}^{\infty} |g_\ell(t)|^2 dt, \qquad (23)$$

or equivalently,

$$\overline{\sigma}^2 = \frac{\sigma_s^2}{\Delta\hat{t}} \sum_{\ell=1}^{L} \int_{-B/2}^{B/2} |G_\ell(f)|^2 df. \qquad (24)$$

If $\overline{\sigma}^2$ is unbounded, the multichannel sampling problem is said to be ill posed. The sufficient and necessary condition for the multichannel sampling problem to be well posed is that the interpolation function is square-integrable; that is, $G_\ell(f)$, $f \in (-B/2, B/2)$ belongs to $L_2(-B/2, B/2)$ [9], [10]. Clearly, to verify the condition requires calculation of the interpolation functions by matrix inversion. An alternative sufficient condition depending only on the input filters states that to achieve the maximum data reduction $R = L$, the condition requires the existence of a positive constant β such that $|\det \mathbf{H}(f)| \geq \beta > 0$, for $f \in (B/2 - \hat{B}, B/2)$ [10].



[FIG5] Setup of parallel magnetic resonance imaging with phased array coils.

PARALLEL MRI AS VIEWED FROM MULTICHANNEL SAMPLING THEORY

While parallel MRI was developed independently of the multichannel sampling theory, their close relationship was recognized in [11] and [12]. This section discusses this connection so as to provide signal processing researchers a familiar view of parallel MRI.

MRI USING PHASED ARRAY COILS

The idea of using phased array coils for MRI dates back to the early 1980s, although practical methods for parallel imaging emerged more recently. The early efforts were focused on building array coils with minimal coil-to-coil coupling [13], [14]. They were successfully used for extended-FOV imaging [15], [16]. Using array coils for fast imaging was explored by a number of groups [17]–[24], although the successful efforts by Sodickson et al. [22] and Pruessmann et al. [23] are largely responsible for the widespread application of this powerful technology.

Figure 5 illustrates parallel imaging using phased array coils. Assume that L coils are used for signal reception. The signal received by the lth coil is given by

$$s_\ell(\vec{k}) = \int_{\text{FOV}} \rho(\vec{r})\, B_{c,\ell}(\vec{r})\, e^{-i2\pi\vec{k}\cdot\vec{r}} d\vec{r}, \qquad (25)$$

where $B_{c,\ell}(\vec{r})$ is the receiving sensitivity function of the ℓth coil, for $\ell = 1, 2, \ldots, L$. Equation (25) is the basic data acquisition equation for MRI using phased array coils, which can be derived from (2)–(4) recognizing the fact that the coils now have spatially nonuniform sensitivity. The term *parallel imaging* is often used for this data acquisition method to emphasize the fact that the $s_\ell(\vec{k})$ are acquired simultaneously for $\ell = 1, 2, \ldots, L$, and the term *sensitivity encoding* is used to refer to the spatial encoding effect of the sensitivity function $B_{c,\ell}(\vec{r})$ in (25). Clearly, a number of things are possible with (25), including extended-FOV imaging and fast imaging with sparse sampling of *k*-space. The latter is the focus of modern parallel imaging research and will be discussed next using multichannel sampling theory.

PARALLEL MRI WITH SPARSE SAMPLING OF K-SPACE

Parallel MRI can be viewed as an application of the multichannel sampling theory by comparing (25) with (18). This connection was originally made in [11] and [12]. In making this connection, the acquired *k*-space signal $s_\ell(\vec{k})$ in parallel MRI corresponds to the output signal $s_\ell(\vec{t})$ of a filter in multichannel sampling, the desired image $\rho(\vec{r})$ corresponds to $S(\vec{f})$, and the coil sensitivity function $B_{c,\ell}(\vec{r})$ corresponds to the filter frequency response $H_\ell(\vec{f})$. The above correspondence is summarized in Table 1.

As described in the introduction, *k*-space can be covered in a number of ways in MRI. While Nyquist criterion has to be satisfied in conventional imaging, the multichannel sampling theory enables sparser sampling of k-space without aliasing artifacts. In the popular case of Cartesian sampling, downsampling can be applied to one direction: the phase encoding direction, and the maximum undersampling factor is L (the number of receiving channels). In this case, the 1-D Papoulis' formulas can be applied directly for image reconstruction due to the separability of Cartesian sampling. However, when undersampling is applied along multiple directions on a non-Cartesian grid, the reconstruction formulas for multichannel sampling may not apply directly. If the sampling pattern has a lattice structure (e.g., zig-zag sampling), the multidimensional, multichannel sampling theory can be used. For other non-Cartesian trajectories used in parallel MRI (such as radial and spiral trajectories), the multidimensional sampling theorem does not provide a well-formulated solution. The reconstruction issue is examined in the section "Discussions."

It is worthwhile to point out that while the analysis/synthesis filters in a filter bank can be designed to satisfy certain "perfect" reconstruction conditions, there is little flexibility in reshaping the coil sensitivity functions in parallel MRI for optimal performance as they are determined by the physical properties of a given receiver system. In practice, the sensitivity functions of receiver coils are not known a priori (as they may also be affected by the object due to the loading effect). Currently, there are two approaches for estimating the sensitivity functions. One approach is to perform a reference scan, from which the sensitivity functions are determined [23]. Another approach is to "densely" sample (satisfying the Nyquist rate) the central k-space to provide self-calibration data [25]. A detailed discussion of the issue can be found in [26].

IMAGE RECONSTRUCTION FROM MULTICHANNEL, UNDERSAMPLED DATA

The multichannel sampling theory provides two approaches for image reconstruction from multichannel k-space data on a Cartesian grid with 1-D undersampling.

K-SPACE METHODS

Based on (6), the k-space signal can be exactly reconstructed by

$$s(k) = \sum_{\ell=1}^{L} \sum_{m=-\infty}^{\infty} s_\ell(m\Delta\hat{k})\, g_\ell(k - m\Delta\hat{k}), \qquad (26)$$

if the coil sensitivity vectors $\{[B_{c,\ell}(x), B_{c,\ell}(x-\hat{B}), \cdots, B_{c,\ell}(x-(L-1)\hat{B})]\}_{\ell=1}^{L}$ are linearly independent for $x \in (B/2-\hat{B}, B/2)$. Or equivalently, the Nyquist samples $s(n\Delta k)$ can be exactly reconstructed using the filter bank formula

$$s(n\Delta k) = \sum_{\ell=1}^{L} \sum_{m=-\infty}^{\infty} s_\ell(mL\Delta k)\, g_\ell((n-mL)\,\Delta k). \qquad (27)$$

The filters $g_\ell(k)$ and $g_\ell(n\Delta k)$ are defined in the same way as in (7) and (17), respectively, except for the notation change (e.g., from k to t).

While (27) gives perfect reconstruction, it is impractical because of the infinite-length filtering required. All practical k-space algorithms use a set of finite-length filters $\tilde{g}_\ell(n\Delta k)$, and the corresponding k-space reconstruction formula becomes

[TABLE 1] EQUIVALENT NOTATIONS IN MULTICHANNEL GENERALIZED SAMPLING AND PARALLEL MRI.

MULTICHANNEL SAMPLING	PARALLEL MRI
TIME DOMAIN ($\vec{t}$)	k-SPACE ($\vec{k}$)
FREQUENCY DOMAIN ($\vec{f}$) B-BANDLIMITED	IMAGE DOMAIN ($\vec{r}$) B-LIMITED FOV
DESIRED SIGNAL $s(\vec{t})$	DESIRED IMAGE IN k-SPACE $s(\vec{k})$
FILTER OUTPUT $S_\ell(\vec{t})$	ACQUIRED k-SPACE SIGNAL $S_\ell(\vec{k})$
SPECTRUM OF DESIRED SIGNAL $S(\vec{f})$	DESIRED IMAGE $\rho(\vec{r})$
LINEAR SYSTEM RESPONSES $H_\ell(\vec{f})$	SPATIAL COIL SENSITIVITIES $B_{c,\ell}(\vec{r})$
INTERPOLATION FUNCTIONS $g_\ell(\vec{f})$	RECONSTRUCTION FILTERS $g_\ell(\vec{k})$
SAMPLING INTERVAL $\Delta\hat{t}$	k-SPACE SAMPLING INTERVAL $\Delta\hat{k}$

$$s(n\Delta k) = \sum_{\ell=1}^{L} \sum_{m=\lfloor n/L\rfloor-N_1}^{\lfloor n/L\rfloor+N_2} s_\ell(mL\Delta k)\, \tilde{g}_\ell((n-mL)\Delta k), \qquad (28)$$

where N_1 and N_2 are integers and $\lfloor\cdot\rfloor$ denotes the floor. A number of methods have been proposed for determining $\tilde{g}_\ell(n\Delta k)$ and implementing (28) (e.g., [27]–[33]), which can be viewed as variants of the well-known simultaneous acquisitions of spatial harmonics (SMASH) algorithm [22] and the generalized autocalibrating partially parallel acquisitions (GRAPPA) algorithm [34].

The SMASH algorithm can be viewed as an approximate implementation of the filter bank reconstruction where the length of the synthesis filter $\tilde{g}_\ell(r\Delta k)$ is assumed to be the same as the undersampling factor. The Nyquist k-space data of the desired image are reconstructed by

$$s[(mL+r)\Delta k] = \sum_{\ell=1}^{L} s_\ell(mL\Delta k)\tilde{g}_\ell(r\Delta k), \quad r=0,\ldots,L-1. \qquad (29)$$

The filter coefficients $\tilde{g}_\ell(r\Delta k)$ are obtained by sampling the analog filters $\tilde{g}_\ell(k)$ at $r\Delta k$, where the $\tilde{g}_\ell(k)$ are the zero-order $(m=0)$ approximation of the perfect reconstruction filters in (7). According to (7), the analog filters are given by

$$\sum_{\ell=1}^{L} \tilde{g}_\ell(k)\, H_\ell(x) \;= e^{i2\pi kx}. \qquad (30)$$

Several extensions of the basic SMASH algorithm have been proposed. For example, AUTO-SMASH [27] and VD-AUTO-SMASH [29] acquire some auto-calibration data to avoid explicit use of the coil sensitivity functions, tailored SMASH [28] allows higher-orders approximation of the perfect reconstruction filters, and generalized SMASH uses finite-length filters to approximate the sensitivity functions [30].

The GRAPPA algorithm [34] also implements (28) with finite-length interpolation filters, but with a key innovation to remove the need for coil sensitivity estimation. Specifically, it expresses the signal from each channel $s_i(n\Delta k)$ as

$$s_i(n\Delta k) = \sum_{\ell=1}^{L} \sum_{m=\lfloor n/L \rfloor - N_1}^{\lfloor n/L \rfloor + N_2} s_\ell(mL\Delta k)\, \tilde{g}_\ell((n-mL)\Delta k). \qquad (31)$$

To determine the interpolation filter coefficients $\tilde{g}_\ell(n\Delta k)$, some calibration k-space data are acquired at the Nyquist rate for each channel. After the $\tilde{g}_\ell(n\Delta k)$ are determined from the calibration data, (31) is used to interpolate $s_i(n\Delta k)$. The final reconstruction is obtained by combining the reconstructions from the interpolated data using a "sum of squares" method [15]. Several extensions of the GRAPPA algorithm have been proposed. For example, use of the GRAPPA operator formalism [31] avoids the need for additional calibration data, cross-validated GRAPPA [32] optimizes the practical filter length, and infinite impulse response (IIR) GRAPPA [33] uses IIR filters in replace of the finite impulse response (FIR) filters.

IMAGE-DOMAIN METHODS

Equation (13) provides a formula for image reconstruction in the image domain

$$\vec{\rho}(x) = \Delta\hat{k}\mathbf{H}^{-T}(x)\, \vec{\rho}^{a}(x). \qquad (32)$$

In practice, with finite sampling, the elements of $\vec{\rho}^{a}(x)$ are determined by

$$\rho_\ell^a(x) = \sum_{m=0}^{N_k-1} s_\ell(m\Delta\hat{k})\, e^{-i2\pi m\Delta\hat{k}x}, \qquad (33)$$

where N_k is the number of undersampled k-space data acquired at each channel.

The sensitivity encoding for fast MRI (SENSE) algorithm [23] is perhaps the most popular image-domain reconstruction method. The basic Cartesian SENSE algorithm implements (32) closely. Many extensions have also been proposed to address various practical issues. For example, regularization has been used to reduce noise amplifications [35], and joint estimation has been used to reduce modeling errors [36].

DISCUSSIONS

While the multichannel sampling theory guarantees "perfect" reconstruction from undersampled, multichannel data under "ideal" conditions, there are several issues that can affect the practical implementation of the theory and the resulting reconstruction quality. This section presents a brief discussion of these issues.

DATA TRUNCATION

The reconstruction formulas for multichannel sampling were derived based on infinite sampling. In practice, finite sampling is used, and a question arises as to how data truncation affects the reconstruction from undersampled, multichannel data. The question was addressed in [37] and, interestingly, the data truncation effects manifest themselves similarly in both conventional Fourier imaging and multichannel imaging under some mild conditions.

More specifically, assume that the input sequence to the filter bank in Figure 3 is a length-N sequence: $s(n\Delta k)$ for $n = -N/2, -N/2+1, \ldots, N/2-1$. The output from the ℓth channel after downsampling will be a length-M sequence: $s_\ell(n\Delta\hat{k})$ for $n = -M/2, -M/2+1, \ldots, M/2-1$. It was shown in [37] that the truncation effect can be described by

$$\hat{\rho}(x) \approx \rho(x) * h(x), \qquad (34)$$

where

$$h(x) = \Delta k \sum_{n=-N/2}^{N/2-1} e^{i2\pi n\Delta kx} = \Delta k \frac{\sin(\pi N\Delta kx)}{\sin(\pi\Delta kx)} e^{-i\pi\Delta kx}. \qquad (35)$$

Equation (34) was used to analyze the data truncation effects in SENSE reconstruction [37] and the approximation is rather accurate when the data truncation is not severe and the coil sensitivity functions are relatively smooth (see [37] for a more detailed description and discussion).

OVERSAMPLING

While the multichannel sampling theory allows undersampling by a factor L (the number of receiver channels), a lower (possibly noninteger) undersampling factor R with $1 < R < L$ is often used in practice. This is called "oversampling" with respect to the maximum undersampling allowed.

In the "oversampling" case, $\mathbf{H}^T(x)$ in (9) has $\lceil R \rceil$ columns when $\lfloor R \rfloor \hat{B} - B/2 < x < B/2$, or $\lfloor R \rfloor$ columns when $B/2 - \hat{B} < x < \lfloor R \rfloor \hat{B} - B/2$, where $\lceil R \rceil$ denotes the ceiling of R. Equation (9) is underdetermined in this case, and thus there are infinite possible solutions for the reconstruction filters $G_\ell(x)$

$$\vec{G}(x) = \Delta\hat{k}(\mathbf{H}^T(x)\, [\mathbf{H}(x)\, \mathbf{H}^T(x)]^{-1}\vec{e} + \vec{N}_{\mathbf{H}(x)}), \qquad (36)$$

where $\vec{N}_{\mathbf{H}(x)}$ denotes an arbitrary vector in the null space of $\mathbf{H}(x)$. Based on the noise analysis in the section "Well Posedness," among the infinite solutions, the one that minimizes the noise power in reconstruction is given by the minimum-norm solution

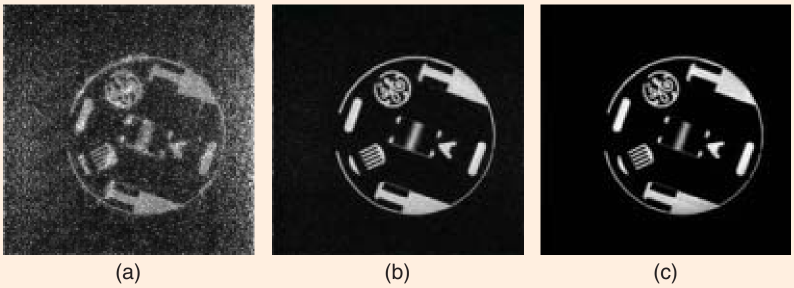
$$\vec{G}(x) = \Delta\hat{k}\mathbf{H}^T(x)\, [\mathbf{H}(x)\, \mathbf{H}^T(x)]^{-1}\vec{e}. \qquad (37)$$

It is easy to see that this oversampled case is usually better conditioned than the critically sampled case because an underdetermined $\mathbf{H}(x)$ is more likely to be well conditioned than a square $\mathbf{H}(x)$. This explains why the noise is improved as the undersampling factor R decreases, as illustrated in Figure 6.

NOISE AMPLIFICATION

Noise amplification is a major practical issue in parallel imaging when the undersampling factors are close to the number of channels. One approach to address the "noise amplification" problem is to design receiver coils with "well-conditioned" sensitivity functions. While it is relatively easy to shape $\mathbf{H}(f)$

in filter bank design, designing phased array coils with "optimal" sensitivity functions for parallel imaging is still an open problem. The noise amplification problem can also be effectively addressed in image reconstruction using, for example, regularization methods (e.g., [35]). Although several regularization methods have been adopted for parallel imaging, much work remains to develop robust methods for regularized image reconstruction from multichannel, undersampled data with characterizable resolution and signal-to-noise ratio performance. This is an area in which signal processing researchers can play a major role.



[FIG6] Reconstruction noise is reduced when the undersampling factor decreases from (a) eight (equal to the number of channels) to (b) four and (c) two.

MODELING ERRORS

In the multichannel sampling theory, it is assumed that we have precise knowledge of the analysis filters. This is not true in parallel MRI as the coil sensitivities need to be estimated from measured data. The effects of using inaccurate insensitivity functions for image reconstruction is dependent on the mathematical condition of $\mathrm{H}(x)$. Specifically, let $\mathrm{H}(x)$ be in error by $\Delta\mathrm{H}(x)$ due to inaccurate sensitivity estimation. The reconstruction in (32) is given by

$$\vec{\rho}(x) + \Delta\vec{\rho}(x) = \Delta\hat{k}[\mathrm{H}(x) + \Delta\mathrm{H}(x)]^{-1}\vec{\rho}^{a}(x), \qquad (38)$$

where $\vec{\rho}(x)$ is the true solution and $\Delta\vec{\rho}(x)$ is the reconstruction error. It can be shown using perturbation analysis [38] that when $\mathrm{H}(x)$ has full rank and $\Delta\mathrm{H}(x)$ has a small L_2-norm, then

$$\frac{\|\Delta\vec{\rho}(x)\|_2}{\|\vec{\rho}(x)\|_2} \le \kappa[\mathrm{H}(x)]\frac{\|\Delta\mathrm{H}(x)\|_2}{\|\mathrm{H}(x)\|_2}, \qquad (39)$$

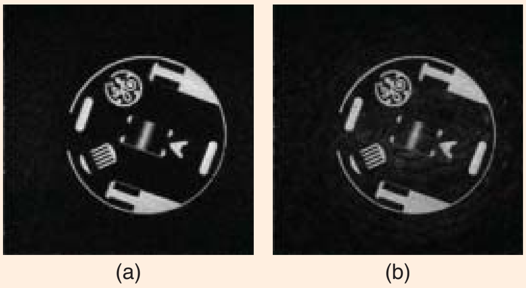
where $\kappa[\mathrm{H}(x)]$ is the condition number of $\mathrm{H}(x)$. Figure 7 gives an example of reconstructions with accurate and inaccurate sensitivities. Because $\mathrm{H}(x)$ is often ill conditioned in practice with large undersampling factors, how to effectively (with performance guarantees) desensitize the reconstruction to modeling errors (and measurement noise) is an important signal processing issue in parallel MRI research.

NON-CARTESIAN SAMPLING TRAJECTORIES

As discussed in the section "Parallel MRI with Sparse Sampling of k-Space," parallel imaging with Cartesian k-space sampling trajectories and 1-D undersampling can be directly mapped to the 1-D multichannel sampling theory. In some MRI experiments, it is desirable to use non-Cartesian sampling trajectories. In this case, if the sampling pattern has a lattice structure as described in the section "Multidimensional Sampling," the multidimensional, multichannel sampling theory can be used. For arbitrary nonuniform sampling, there is no "standard" multichannel sampling theory available, and there are two approaches to handle the problem in existing parallel MRI methods. The first approach is to directly discretize (25) with some ideal voxel functions $\gamma(\vec{r}, \vec{r}_n)$ (e.g., [39]). The image voxels can then be determined by solving a large linear system equation. This approach requires knowledge of the coil sensitivity functions $B_{c,\ell}(\vec{r})$. The second approach is to directly interpolate the k-space data using an interpolating kernel determined from reference or calibrating data, as is done in GRAPPA (e.g., [40] and references cited). While both approaches have worked well, it is clear that much work remains to optimize data acquisition and image reconstruction for parallel MRI with non-Cartesian sampling trajectories.

CONCLUSION

Parallel MRI using phased array coils can be viewed as an application of the multichannel sampling theory. Specifically, in the case of uniform 1-D undersampling, Papoulis' classical reconstruction formulas correspond well to the existing parallel MRI reconstruction algorithms, and a number of practical issues can be analyzed in this context. However, parallel MRI also presents several unique signal processing problems, whose solutions can help maximize the potential of parallel MRI for fast imaging. While existing parallel MRI methods were developed independently of the multichannel sampling theory, making such a connection



[FIG7] SENSE reconstructions with sensitivities estimated using (a) high-resolution and (b) low-resolution reference scans. Eight channels and an undersampling factor of four were used.

may help develop more optimal methods for parallel MRI data acquisition and image reconstruction.

ACKNOWLEDGMENTS

The work was supported in part by NSF CBET-0731226 and CBET-0730623. The authors would also like to thank Justin Haldar for useful comments and Dong Liang and Yuchou Chang for preparing the figures.

AUTHORS

Leslie Ying (leiying@uwm.edu) received her B.E. degree in electronics engineering from Tsinghua University, China, in 1997 and her M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1999 and 2003, respectively. She is now an associate professor of electrical engineering and computer science at the University of Wisconsin-Milwaukee. Her research interests include sampling and interpolation, image reconstruction, and MRI. She received the CAREER Award from the National Science Foundation in 2009.

Zhi-Pei Liang (z-liang@illinois.edu) is a professor of electrical and computer engineering and cochair of the Integrative Imaging Theme of the Beckman Institute at the University of Illinois, Urbana-Champaign. His research interests include MRI and spectroscopy, image-formation theory, algorithms, and applications.

REFERENCES

[1] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[2] Z.-P. Liang, “Spatiotemporal imaging with partially separable functions,” in *Proc. IEEE Int. Symp. Biomedical Imaging*, 2007, pp. 988–991.

[3] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[4] A. Papoulis, “Generalized sampling expansion,” *IEEE Trans. Circuits Syst.*, vol. CAS-24, no. 11, pp. 652–654, 1977.

[5] J. L. Brown, Jr., “Multi-channel sampling of low-pass signals,” *IEEE Trans. Circuits Syst.*, vol. CAS-28, no. 2, pp. 101–106, 1981.

[6] P. P. Vaidyanathan, “Classical sampling theorems in the context of multirate and polyphase digital filter bank structures,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 9, pp. 1480–1495, 1988.

[7] P. P. Vaidyanathan, “Theory and design of M-channel maximally decimated quadrature mirror filters with arbitrary M, having the perfect-reconstruction property,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 4, pp. 476–492, 1987.

[8] K. F. Cheung, “A multidimensional extension of Papoulis’ generalized sampling expansion with application in minimum density sampling,” in *Advanced Topics in Shannon Sampling and Interpolation Theory*, R. J. Marks, II, Ed. Berlin, Germany: Springer-Verlag, 1993, pp. 86–119.

[9] K. F. Cheung and R. J. Marks, II, “Ill-posed sampling theorems,” *IEEE Trans. Circuits Syst.*, vol. CAS-32, no. 5, pp. 481–484, 1985.

[10] J. L. Brown, Jr. and S. D. Cabrera, “On well-posedness of the Papoulis generalized sampling expansion,” *IEEE Trans. Circuits Syst.*, vol. 38, no. 5, pp. 554–556, 1991.

[11] M. Drobnitzky, “Image reconstruction via generalized sampling expansion in parallel MR acquisition techniques,” in *Proc. First Würsburg Workshop Parallel Imaging*, 2001, p. 96.

[12] Z.-P. Liang, L. Ying, D. Xu, and L. Yuan, “Parallel imaging: Some signal processing issues and solutions,” in *Proc. IEEE Int. Symp. Biomedical Imaging*, 2004, pp. 1204–1207.

[13] J. F. Schenck, H. R. Hart, T. H. Foster, W. A. Edelstein, P. A. Bottomley, C. J. Hardy, R. A. Zimmerman, and L. T. Bilaniuk, “Improved MR imaging of the orbit 1.5T with surface coils,” *Amer. J. Neuroradiol.*, vol. 6, no. 5, pp. 193–196, 1985.

[14] J. S. Hyde, A. Jesmanowicz, T. M. Gristand, W. Froncisz, and J. B. Kneeland, “Quadrature detection surface coil,” *Magn. Reson. Med.*, vol. 4, no. 2, pp. 179–184, 1987.

[15] P. B. Roemer, W. A. Edelstein, C. E. Hayes, S. P. Souza, and O. M. Mueller, “The NMR phased array,” *Magn. Reson. Med.*, vol. 16, no. 2, pp. 192–225, 1990.

[16] C. E. Hayes, N. Hattes, and P. B. Roemer, “Volume imaging with MR phased arrays,” *Magn. Reson. Med.*, vol. 18, no. 2, pp. 309–319, 1991.

[17] M. Hutchinson and U. Raff, “Fast MRI data acquisition using multiple detectors,” *Magn. Reson. Med.*, vol. 6, no. 1, pp. 87–91, 1988.

[18] D. Kwiat, S. Einav, and G. Navon, “A decoupled coil detector array for fast image acquisition in magnetic resonance imaging,” *Med. Phys.*, vol. 18, no. 2, pp. 251–265, 1991.

[19] J. B. Ra and C. Y. Rim, “Fast imaging using subencoding data sets from multiple detectors,” *Magn. Reson. Med.*, vol. 30, no. 1, pp. 142–145, 1993.

[20] J. W. Carlson and T. Minemura, “Imaging time reduction through multiple receiver coil data acquisition and image reconstruction,” *Magn. Reson. Med.*, vol. 29, no. 5, pp. 681–687, 1993.

[21] J. R. Kelton, R. L. Magin, and S. M. Wright, “An algorithm for rapid image acquisition using multiple receiver coils,” in *Proc. Int. Society for Magnetic Resonance in Medicine*, 1989, p. 1172.

[22] D. K. Sodickson and W. J. Manning, “Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays,” *Magn. Reson. Med.*, vol. 38, no. 4, pp. 591–603, Oct. 1997.

[23] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Böesiger, “SENSE: Sensitivity encoding for fast MRI,” *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, 1999.

[24] M. P. McDougall and S. M. Wright, “64-channel array coil for single echo acquisition magnetic resonance imaging,” *Magn. Reson. Med.*, vol. 54, no. 2, pp. 386–392, 2005.

[25] C. A. McKenzie, E. N. Yeh, M. A. Ohliger, M. D. Price, and D. K. Sodickson, “Self-calibrating parallel imaging with automatic coil sensitivity extraction,” *Magn. Reson. Med.*, vol. 47, no. 3, pp. 529–538, 2002.

[26] M. Blaimer, F. Breuer, M. Mueller, R. M. Heidemann, M. A. Griswold, and P. M. Jakob, “SMASH, SENSE, PILS, GRAPPA: How to choose the optimal method,” *Top. Magn. Reson. Imaging*, vol. 15, no. 4, pp. 223–236, 2004.

[27] P. M. Jakob, M. A. Griswold, R. R. Edelman, and D. K. Sodickson, “AUTO-SMASH: A self-calibrating technique for SMASH imaging,” *MAGMA*, vol. 7, no. 1, pp. 42–54, 1998.

[28] D. K. Sodickson, “Tailored SMASH image reconstructions for robust in vivo parallel MR imaging,” *Magn. Reson. Med.*, vol. 44, no. 2, pp. 243–245, 2000.

[29] R. M. Heidemann, M. A. Griswold, A. Haase, and P. M. Jakob, “VD-AUTO-SMASH imaging,” *Magn. Reson. Med.*, vol. 45, no. 6, pp. 1066–1074, 2001.

[30] M. Bydder, D. J. Larkman, and J. V. Hajnal, “Generalized SMASH imaging,” *Magn. Reson. Med.*, vol. 47, no. 1, pp. 160–170, 2002.

[31] M. A. Griswold, M. Blaimer, F. Breuer, R. M. Heidemann, M. Mueller, and P. M. Jakob, “Parallel magnetic resonance imaging using the GRAPPA operator formalism,” *Magn. Reson. Med.*, vol. 54, no. 6, pp. 1553–1556, 2005.

[32] R. Nana, T. Zhao, K. Heberlein, S. M. LaConte, and X. Hu, “Cross-validation-based kernel support selection for improved GRAPPA reconstruction,” *Magn. Reson. Med.*, vol. 59, no. 4, pp. 819–825, 2008.

[33] Z. Chen, J. Zhang, R. Yang, P. Kellman, L. A. Johnston, and G. F. Egan, “IIR GRAPPA for parallel MR image reconstruction,” *Magn. Reson. Med.*, vol. 63, no. 2, pp. 502–509, 2010.

[34] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, “Generalized autocalibrating partially parallel acquisitions (GRAPPA),” *Magn. Reson. Med.*, vol. 47, no. 6, pp. 1202–1210, 2002.

[35] Z. P. Liang, R. Bammer, J. Ji, N. Pelc, and G. Glover, “Making better SENSE: Wavelet denoising, Tikhonov regularization, and total least squares,” in *Proc. Int. Society for Magnetic Resonance in Medicine*, 2002, p. 2388.

[36] L. Ying and J. Sheng, “Joint image reconstruction and sensitivity estimation in SENSE (JSENSE),” *Magn. Reson. Med.*, vol. 57, no. 6, pp. 1196–1202, 2007.

[37] L. Yuan, L. Ying, D. Xu, and Z.-P. Liang, “Truncation effects in SENSE reconstruction,” *Magn. Reson. Imaging*, vol. 24, no. 10, pp. 1311–1318, 2006.

[38] G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[39] K. P. Pruessmann, M. Weiger, and P. Böernert, “Advances in sensitivity encoding with arbitrary *k*-space trajectories,” *Magn. Reson. Med.*, vol. 46, no. 4, pp. 638–651, 2001.

[40] N. Seiberlich, F. A. Breuer, R. M. Heidemann, M. Blaimer, M. A. Griswold, and P. M. Jakob, “Reconstruction of undersampled non-Cartesian data sets using pseudo-Cartesian GRAPPA in conjunction with GROG,” *Magn. Reson. Med.*, vol. 59, no. 5, pp. 1127–1137, 2008.

[SP]

William Roberts, Hao He, Jian Li, and Petre Stoica

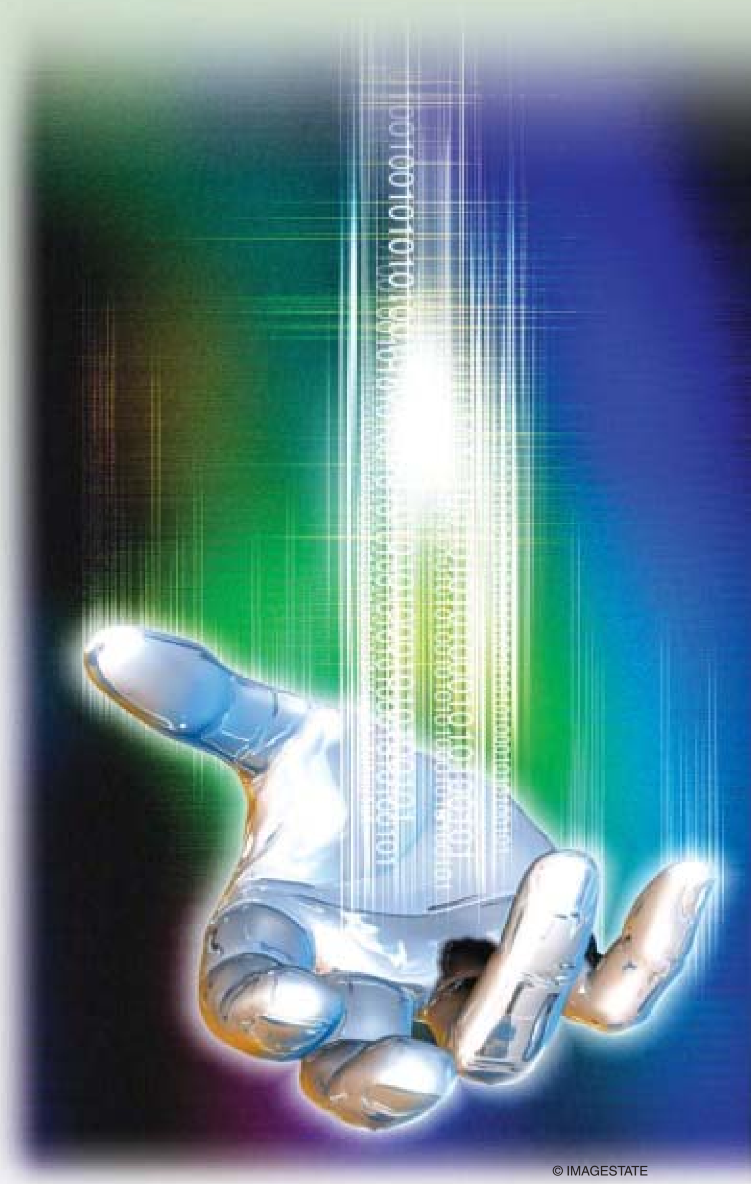
Probing Waveform Synthesis and Receiver Filter Design

A review of recent novel, cyclic approaches to single and multiple waveform designs

Probing waveform synthesis and receive filter design play crucial roles in achievable performance for many active sensing applications, including radar, sonar, medical imaging, and communications (channel estimation and spread spectrum). A flexible receive filter design approach can be used to compensate for missing features of the probing waveforms, at the costs of lower signal-to-noise ratio (SNR) and higher computational complexity. A well-synthesized waveform, meaning one with good auto- and cross-correlation properties, can reduce computational burden at the receiver and improve performance. In this article, we will highlight the interplay between waveform synthesis and receiver design. We will provide a tutorial review of recent novel, cyclic approaches to single and multiple waveform designs. Both aperiodic and periodic correlations will be considered. We show that by making use of fast Fourier transforms (FFTs), we can now efficiently design sequences that were previously impossible to synthesize. Furthermore, we will provide an overview of some advanced techniques for receiver design, including data-independent instrumental variables (IV) filters and a data-adaptive iterative approach. We will show how these designs can significantly outperform conventional techniques in various active sensing applications.

INTRODUCTION

Areas of active sensing (including radar, sonar, medical imaging, and communications) have garnered the attention of researchers for decades. The goal of any active sensing application is the transmission and reception of one or more chosen waveforms. A received signal may be analyzed to determine properties of a propagation medium, as in channel estimation for communications, or to estimate the location and strength of targets in a scene, as in medical imaging for breast cancer detection.



© IMAGESTATE

Digital Object Identifier 10.1109/MSP.2010.936724

1053-5888/10/$26.00©2010IEEE

Performance, in a most general sense, is measured simply by the accuracy with which an interpretation of the received signal matches the true information in the scene.

Not surprisingly, a system's performance is linked directly to the behavior of its transmitted waveform(s), which has resulted in a vast amount of literature devoted to the study of sequence design (e.g., see [1]–[4]). The foundation of research in signal transmission has been built, over the last century, on work from such notable theorists as Shannon, Nyquist, and Tesla. Following World War II, the development and characterization of transmission sequences was furthered by such researchers as Woodward, Barker, Frank, and Golomb. Waveform synthesis remains a dynamic research area even in modern times, as continued improvements in waveform generation and receiving hardware allow for increasingly advanced sequence design. As computational limitations on waveform design and transmission are relaxed, new approaches and improved performance become possible.

In this article, we will offer a tutorial description of some of the very latest algorithms for sequence design. Our discussion is an extension to the recent special issue [5], which focused on several new strategies for waveform development in radar systems. In [6], a design scheme for periodic constant amplitude with zero autocorrelation (CAZAC) sequences is presented for single waveform transmissions. Another article [7] in the special issue explores the design of complementary sequences. Herein, we focus primarily on waveforms with good aperiodic correlations, and we show that our designs can be extended to design periodic CAZAC sequences as well. We consider both single sequence and multiple sequence set designs. The design of multiple sequences with good auto- and cross-correlation properties is demanded in the emerging fields of multiple-input, multiple-output (MIMO) radar and MIMO communications (see, e.g., [8]–[11]). Further, since our goal is improved system performance, we will give an overview of several novel receiver design techniques. We will clearly illustrate how receiver design can be used to compensate for deficiencies in waveform synthesis.

THE ACTIVE SENSING PROBLEM

Consider the synthesis of a single sequence with good aperiodic autocorrelation, which is widely needed in conventional single-input, single-output (SISO) radar, sonar, and medical imaging applications. In SISO radar, for example, a pulse is transmitted in the direction of a scene of interest [12]. The signal is reflected by targets in the scene, which could be at different angular and range locations relative to the radar. The reflected signals, which are attenuated and time-shifted versions of the transmitted waveform, are linearly combined at the receive antenna (which could be the same as or different from the one used for transmission). Signal processing of the received signal is performed to determine unknown properties of the targets, such as their range, radar cross section (RCS), and speed (or Doppler shift).

We let $s(t)$ denote a transmitted waveform of duration τ and comprising N subpulses (so that each subpulse has a duration of τ/N). We can then represent $s(t)$ by the vector $\mathbf{x}$, whose components correspond to the phase-coded amplitude of each subpulse (we assume rectangular subpulses)

$$\mathbf{x} = [x(1), \quad x(2), \quad \ldots, \quad x(N)]^T, \tag{1}$$

where $(\cdot)^T$ denotes the transpose operation. Due to hardware constraints (such as the limitations of the power amplifier) in practice, components of the transmitted waveform are commonly restricted to being constant modulus. Without loss of generality, we consider $\{x(n)\}_{n=1}^{N}$ being unimodular, so that

$$x(n) = e^{j\phi_n}, \quad n = 1, \ldots, N, \tag{2}$$

where ϕ_n represents the phase of $x(n)$. If the set of targets in the scene are represented by their RCSs $\{\kappa_{r,l}\}$ (for $r = 1, \ldots, R$ denoting the range bin and $l = 1, \ldots, L$ denoting the Doppler bin of a target), then the received signal $\mathbf{y}_{r'}$ (aligned with the transmitted waveform's reflection from a range bin of interest r') can be modeled as

$$\mathbf{y}_{r'} = \kappa_{r',l'}\tilde{\mathbf{x}}_{l'} + \sum_{\substack{n=-N+1 \\ (r'+n,l)\neq(r',l')}}^{N-1} \sum_{l=1}^{L} \kappa_{r'+n,l}\, \mathbf{J}_n \tilde{\mathbf{x}}_l + \boldsymbol{\varepsilon}, \tag{3}$$

where $\kappa_{r',l'}$ refers to the reflection coefficient of a target of interest and $\boldsymbol{\varepsilon}$ refers to the noise component of the received signal. Further, $\mathbf{J}_n$ is a shift matrix designed to temporally align the reflected signal from a target that lies n range bins away from the bin of interest

$$\mathbf{J}_n = \begin{bmatrix} 0 & & & & 0 \\ \vdots & & & & \\ 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ 0 & & & \underbrace{1 \cdots 0}_{n+1} & \end{bmatrix}_{N\times N} = \mathbf{J}_{-n}^T, \quad n = 0, 1, \ldots, N-1. \tag{4}$$

We assume that, in general, $\kappa_{r,l} = 0$ for any r such that $r \notin \{1, \ldots, R\}$. We let $\tilde{\mathbf{x}}_l = \mathbf{x} \odot \mathbf{a}_l$ denote the Doppler shifted waveform ($\odot$ refers to the Hadamard product operation), and

$$\mathbf{a}_l = [1, \quad e^{j\omega_l}, \quad \ldots, \quad e^{j\omega_l(N-1)}]^T, \quad l = 1, \ldots, L, \tag{5}$$

where ω_l represents the Doppler frequency for the lth Doppler bin (we assume L bins divide the Doppler interval of interest). We illustrate the model for $\mathbf{y}_{r'}$ in Figure 1 (the Doppler effect is not shown). The problem of interest for this SISO radar case, as well as for any other active sensing application, is the successful estimation of the unknown target coefficients given by $\{\kappa_{r,l}\}$.

TRANSMIT WAVEFORM DESIGN

Neglecting interferences from other range bins, which are represented by the clutter returns in Figure 1, a matched filter can be used to provide optimal performance (the highest SNR) in the presence of stochastic additive white noise. When clutter affects the received signal, a matched filter will function optimally only if the autocorrelation sidelobe terms of the transmit waveform, given by

$$r(\Delta t) \triangleq \int s(t)s^*(t-\Delta t)dt, \quad \forall \Delta t \neq 0 \tag{6}$$

are zero, where $(\cdot)^*$ denotes the complex conjugate for scalars and the conjugate transpose operation for vectors and matrices.

For most modern sensing systems, filtering at the receiver is performed digitally. Further, when rectangular subpulses are adopted at the transmitter, values of $r(\Delta t)$ can be obtained exactly by a linear combination of two neighboring correlations evaluated at integer multiples of τ/N (approximations can be made when nonrectangular shaping pulses are used) [3]. Thus, in most practical cases, we can restrict our attention to the autocorrelation of the discrete sequence $\{x(n)\}_{n=1}^{N}$

$$r_k = \sum_{n=k+1}^{N} x(n)x^*(n-k) = r_{-k}^*, \quad k = 0, \ldots, N-1. \tag{7}$$

No signal in practice can have zero sidelobes for all $k \neq 0$ in (7) (since, e.g., $|r_{N-1}| = |r_{-N+1}| = 1$ for all unimodular sequences). Therefore, a legitimate goal in transmit sequence design would be to construct $\{x(n)\}_{n=1}^{N}$ such that the autocorrelation sidelobes $\{r_k, (k \neq 0)\}$ are as small as possible. In other words, waveforms with high merit factors (MFs) are desirable [13], [14], where we let

$$\text{MF} = \frac{N^2}{\text{ISL}}, \tag{8}$$

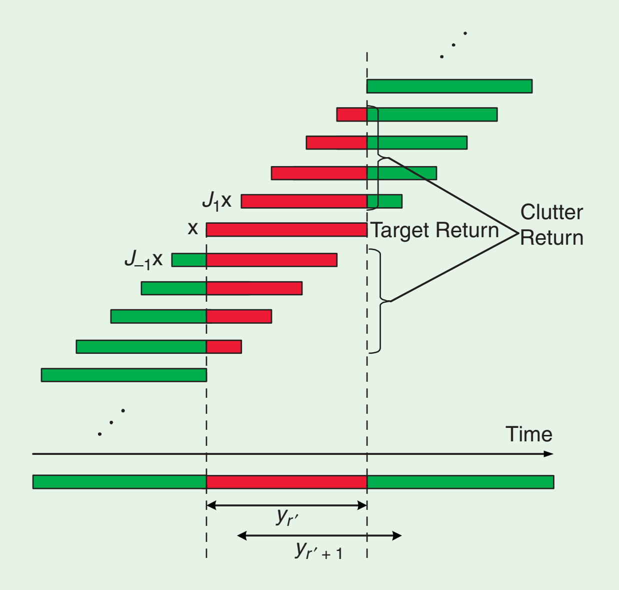
with

$$\text{ISL} = \sum_{\substack{k=-(N-1)\\ k\neq 0}}^{N-1} |r_k|^2. \tag{9}$$

ISL refers to the integrated sidelobe level of the autocorrelation function. In the next subsection, we will review several existing waveforms (including several phase-coded waveforms) that have good autocorrelation properties.

A REVIEW OF EXISTING WAVEFORMS

In 1953, Barker introduced a set of binary codes (meaning $\phi_n \in \{-\pi, \pi\}$ for $n = 1, \ldots, N$) that yield a peak-to-peak sidelobe ratio of N and subsequently the highest MF for binary sequences of equal length [15]. However, the longest known Barker sequence is of length 13, and researchers contend that no longer waveforms can satisfy the Barker criteria [3], [16], [17]. To identify binary sequences that yield a maximum MF (for a given N) requires an exhaustive search whose computational complexity increases exponentially with the length, which quickly proves intractable as N increases. At the cost of increased hardware complexity, the binary restriction can be relaxed to design unimodular sequences (which may or may not use a finite alphabet) with lower sidelobe levels and higher merit factors.



[FIG1] Received signal aligned with the return from a target in range bin r'.

Many well-known unimodular phase-coded signals are derived from the phase history of a chirp waveform. A chirp is a linear frequency-modulated (LFM) pulse whose frequency is swept linearly over a bandwidth B in the sequence's time duration τ. Chirp waveforms have been widely used for radar applications since World War II, as they possess relatively low peak sidelobe levels and are mostly tolerant to shifts in Doppler frequency [3]. In addition, chirp signals have spectral efficiency, meaning the power of the waveforms is dispersed evenly throughout the frequency spectrum, which allows for high range resolution.

The chirp waveform $s(t)$ is given by

$$s(t) = \frac{1}{\sqrt{\tau}} e^{j\pi \frac{B}{\tau} t^2}, \quad 0 \leq t \leq \tau, \tag{10}$$

where B/τ is the chirp rate of the signal. By sampling $s(t)$ at time intervals $t_s(n) = n/B$, for $n = 1, \ldots, N$ $(N = BT)$, and by omitting the multiplicative term $1/\sqrt{\tau}$, the following discrete sequence is obtained:

$$\begin{aligned} x(n) = s(t_s(n)) &= e^{j\pi \frac{B}{T}\left(\frac{n}{B}\right)^2} \\ &= e^{j\pi \frac{n^2}{BT}} = e^{j\pi \frac{n^2}{N}}, \quad n = 1, \ldots, N. \end{aligned} \tag{11}$$

The signal $\{x(n)\}_{n=1}^{N}$ shown in (11) has perfect periodic autocorrelation if N is even, meaning that all periodic autocorrelation sidelobes are zero

$$\tilde{r}(k) = \sum_{n=1}^{N} x(n)x^*((n+k) \bmod N) = 0, \quad 1 \leq k \leq N-1, \tag{12}$$

where $(a \bmod b) = a - \lfloor a/b \rfloor b$. We refer to waveforms with perfect periodic autocorrelations as CAZAC sequences, which were thoroughly reviewed in [6]. A set of CAZAC sequences can also be constructed, for odd values of N, by altering (11) as follows:

$$x(n) = e^{j\pi\frac{n(n-1)}{N}}, \ \ n = 1, \dots, N, \tag{13}$$

which is the famous Golomb sequence [18].

The Frank code [19], which was first presented in 1963, is perhaps the most well-known CAZAC sequence. Frank signals are also derived from the phase history of a chirp waveform, and are defined for a square $N = K^2$ length sequence as

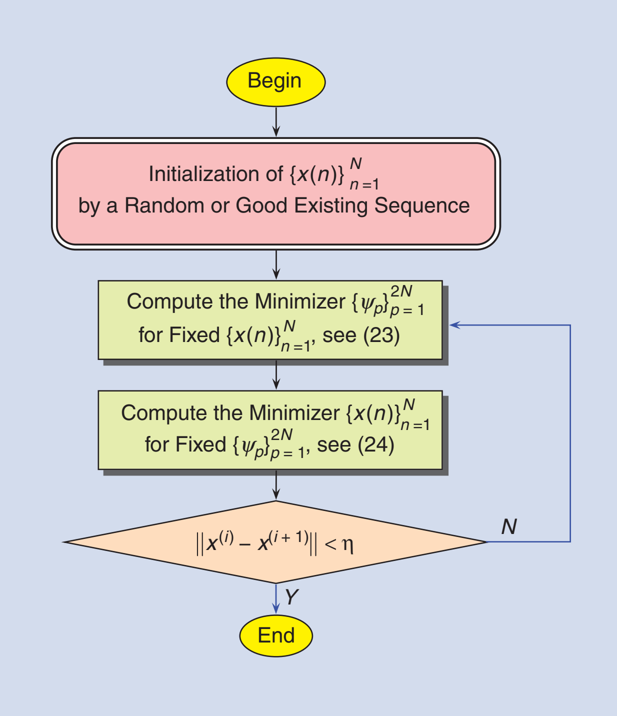
$$x((m-1)K + p) = e^{j2\pi\frac{(m-1)(p-1)}{K}}, \ \ m, p = 1, \dots, K. \tag{14}$$

Similarly, P4 sequences [20] are phase-coded CAZAC waveforms whose phases are quadratic functions of n. The P4 sequence is defined for any length N as

$$x(n) = e^{j\frac{2\pi}{N}(n-1)\left(\frac{n-1-N}{2}\right)}, \ \ n = 1, \dots, N. \tag{15}$$

CAZAC sequences, both chirp like and nonchirp like, have been shown to exist for any length N [6] (infinite number of sequences exist for some N, in fact). In contrast, the design of signals with low ISL levels (high MF values), and thus good aperiodic correlation properties, has proven more challenging to researchers. The need for low ISL waveforms, as opposed to CAZAC sequences, is entirely dependent on the application and directly relates to the stationarity of the scene or channel, the maximum signal delay, and the SNR. We focus herein on the design of signals with low aperiodic correlation levels.

[TABLE 1] FLOWCHART OF THE CAN ALGORITHM.


Begin
Initialization of $\{x(n)\}_{n=1}^{N}$
by a Random or Good Existing Sequence
Compute the Minimizer $\{\psi_p\}_{p=1}^{2N}$
for Fixed $\{x(n)\}_{n=1}^{N}$, see (23)
Compute the Minimizer $\{x(n)\}_{n=1}^{N}$
for Fixed $\{\psi_p\}_{p=1}^{2N}$, see (24)
$\|x^{(i)} - x^{(i+1)}\| < \eta$
N
Y
End

To find sequences with low autocorrelation levels that lack a closed-form expression (unlike the previous signals), researchers have used gradient descent and stochastic optimization techniques (see, e.g., [21]–[23]). These algorithms are usually computationally expensive and only perform well for small values of N, such as $N{\sim}10^2$. For large values of N, a series of recently proposed cyclic algorithms (CAs) can be used to effectively minimize the ISL-related metrics locally. We outline these algorithms in the next few subsections.

THE CAN ALGORITHM

We first present a synopsis of the CA-new (CAN) algorithm proposed in [24]. Recalling the ISL definition offered in (9), and by applying the Parseval equality, the ISL of a sequence can be expressed in the frequency domain as

$$\text{ISL} = \frac{1}{2N}\sum_{p=1}^{2N}[\Phi(\theta_p) - N]^2, \tag{16}$$

where $\{\Phi(\theta_p)\}_{p=1}^{2N}$ is the DFT of $\{r(k)\}_{k=-N+1}^{N-1}$ at frequencies $\{\theta_p = 2\pi p/2N\}_{p=1}^{2N}$ [25]. Since the DFT of $\{r(k)\}_{k=-N+1}^{N-1}$ yields the spectral density function of $\{x(n)\}_{n=1}^{N}$, so that $\Phi(\theta_p) = |\sum_{n=1}^{N} x(n)e^{-j\theta_p n}|^2$, the ISL in (16) can be further expressed as

$$\text{ISL} = \frac{1}{2N}\sum_{p=1}^{2N}\left[\left|\sum_{n=1}^{N} x_n e^{-j\theta_p n}\right|^2 - N\right]^2. \tag{17}$$

Minimization of the ISL metric in (17), which is a quartic function of $\{x(n)\}_{n=1}^{N}$, can prove computationally challenging. To simplify, we instead consider the following "almost equivalent" problem [26], [24]

$$\min_{\{x(n)\}_{n=1}^{N};\{\psi_p\}_{p=1}^{2N}} \sum_{p=1}^{2N}\left|\sum_{n=1}^{N} x_n e^{-j\theta_p n} - \sqrt{N}e^{j\psi_p}\right|^2 \tag{18}$$
$$\text{s.t. } |x(n)| = 1, \ \ n = 1, \dots, N.$$

Sequences that minimize the ISL equation in (17) and those that solve the ISL-related minimization problem in (18) will, in general, be different. However, we contend that a sequence that makes the cost function in (18) small will certainly lead to a small ISL value in (17) (please see [27] for further discussion). To within a multiplicative constant, the criterion in (18) can be rewritten as

$$\|\mathbf{A}^*\mathbf{w} - \mathbf{v}\|^2, \tag{19}$$

where

$$\mathbf{w} = \left[x(1), \ \dots, \ x(N), \ 0, \ \dots, \ 0\right]^T_{2N\times 1}, \tag{20}$$

$$\mathbf{v} = \frac{1}{\sqrt{2}}[e^{j\psi_1}, \ \dots, \ e^{j\psi_{2N}}]^T, \tag{21}$$

and

$$\mathbf{A}^* = \frac{1}{\sqrt{2N}}\begin{bmatrix} e^{-j\theta_1} & \cdots & e^{-j2N\theta_1} \\ \vdots & \ddots & \vdots \\ e^{-j\theta_{2N}} & \cdots & e^{-j2N\theta_{2N}} \end{bmatrix}. \tag{22}$$

If $\mathbf{f} = \mathbf{A}^{*}\mathbf{w}$ represents the Fourier transform of $\mathbf{w}$, then for fixed $\mathbf{f}$, the minimizer of (19) is given by

$$\psi_p = \arg(f(p)), \ \ p = 1, \ldots, 2N. \tag{23}$$

Similarly, for a given $\mathbf{v}$, and if $\mathbf{g} = \mathbf{A}\mathbf{v}$ denotes the inverse Fourier transform of $\mathbf{v}$, then the minimizing sequence $\{x(n)\}_{n=1}^{N}$ of (19) is given by

$$x(n) = e^{j\arg(g(n))}, \ \ n = 1, \ldots, N. \tag{24}$$

The steps of CAN, to provide the cyclic minimization of the ISL-related metric in (18), are summarized in Table 1 (where $\mathbf{x}^{(i)}$ denotes the sequence obtained at the ith iteration). Note that in Table 1, η is a predefined threshold, such as 10^{-3}. Due to its simple FFT operations [see (22)–(24)], CAN can be used to provide waveform synthesis on an ordinary PC for very large values of N, such as $N{\sim}10^6$.

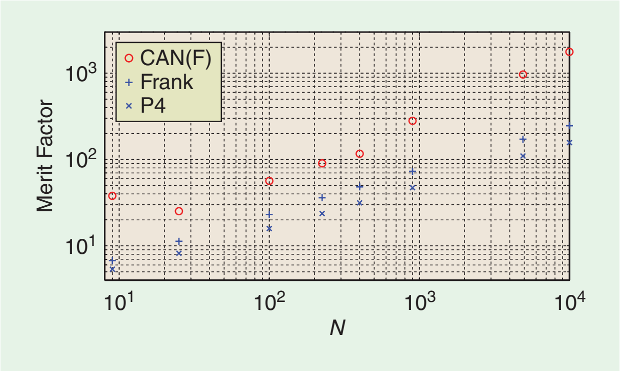
CAN (as well as periodic CAN, which is described in the section "The Periodic Correlation Case") shares a close relationship with the Gerchberg-Saxton algorithm (GSA) [28], which was originally presented in the optics literature more than 35 years ago. In Appendix A, we will review a version of GSA and explain its connection to the cyclic algorithms described in this article.

In Figure 2, we compare the merit factors of the P4, Frank, and CAN (initialized with a Frank sequence) sequences for the following lengths: $N = 3^2, 5^2, 10^2, 15^2, 20^2, 30^2, 70^2$, and 100^2 (note that each N is chosen to be a square to cater to the Frank sequence; the CAN algorithm does not have such a restriction). The results are shown using a log-log scale. The CAN sequence provides the highest merit factor for each value of N considered. When $N = 100^2$, the CAN sequence provides the largest merit factor of 1,769.05, which is several times larger than that given by the Frank sequence (which is 246.39). Although a Frank sequence was used here to initialize CAN, a similar result would have been obtained by initializing the algorithm with a P4 or Golomb sequence (since these chirp-based waveforms are closely related).

We provide the autocorrelation of a Frank sequence and CAN sequence (again initialized with a Frank sequence) for length $N = 100^2$ in Figure 3(a) and (b), respectively. In addition to its lower ISL value, the CAN sequence has a lower PSL (-57.27 dB) compared to the Frank signal (-49.94 dB).

THE CA ALGORITHM

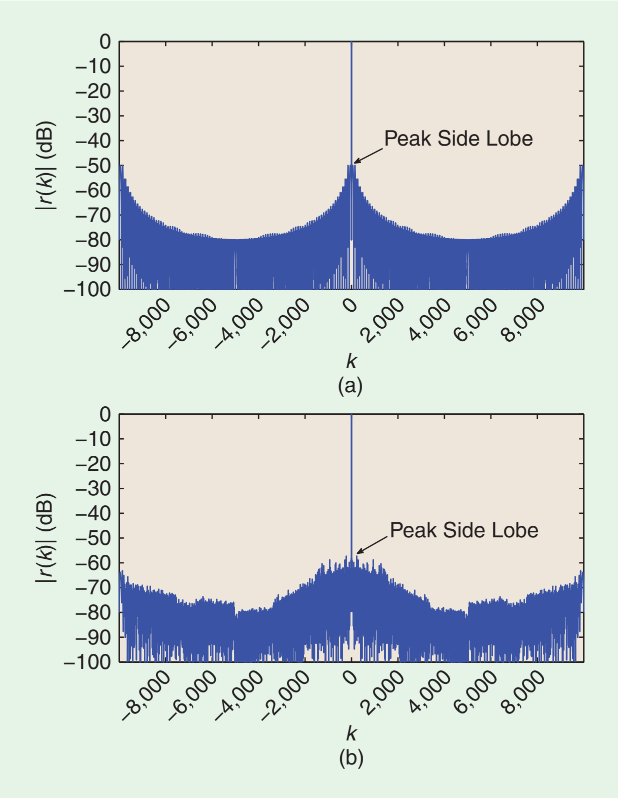
In some cases, the maximum difference between the arrival times of the sequence of interest and of the interference is (much) smaller than the duration of the emitted signal (see, e.g., [29]–[31]). Consequently, for transmit sequence design in such instances, the interest lies in making $\{|r(k)|\}_{k=1}^{P-1}$ small, for some $P < N$, instead of trying to minimize all correlation sidelobes $\{|r(k)|\}_{k=1}^{N-1}$. The value of P is selected based on a priori knowledge about the application. In wireless communications, for example, significant channel tap coefficients can occur only up to a certain known maximum delay (P is chosen as the said delay). In this section, we briefly summarize the CA algorithm, which serves as an extension to CAN for this $P < N$ case. Further details, as well as an application of CA to multiple sequence sets, can be found in [32] and [26].



[FIG2] The merit factor versus the sequence length *N* for P4, Frank and CAN (initialized with a Frank sequence) sequences.

Define the following matrix:

$$\mathbf{X} = \begin{bmatrix} x(1) & & 0 \\ \vdots & \ddots & \\ \vdots & & x(1) \\ x(N) & & \vdots \\ & \ddots & \vdots \\ 0 & & x(N) \end{bmatrix}_{(N+P-1)\times P} . \tag{25}$$



[FIG3] The autocorrelations of (a) a Frank sequence with *N* = 100^2 and (b) a CAN sequence, initialized with a Frank sequence, with *N* = 100^2.

It follows that

$$\mathbf{X}^*\mathbf{X} = \begin{bmatrix} r(0) & r(1)^* & \cdots & r^*(P-1) \\ r(1) & r(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & r^*(1) \\ r(P-1) & \cdots & r(1) & r(0) \end{bmatrix}_{P\times P}. \qquad (26)$$

Minimization of the autocorrelation terms $\{|r(k)|\}_{k=1}^{P-1}$ can be achieved by minimizing the criterion $\|\mathbf{X}^*\mathbf{X} - N\mathbf{I}\|^2$. Similar to (18), we can instead define the following "almost equivalent," computationally feasible, minimization problem:

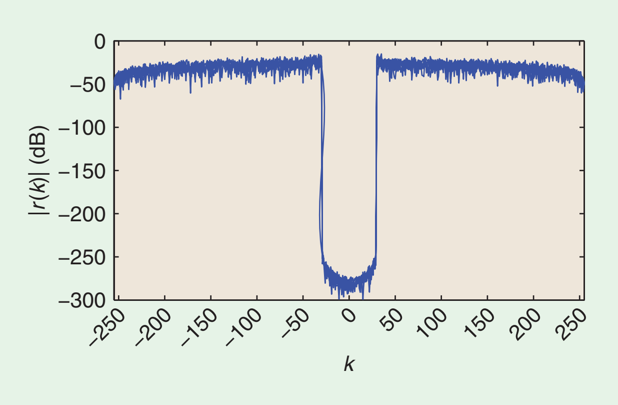
$$\begin{aligned} &\min_{\{x(n)\}_{n=1}^{N};\mathbf{U}} \|\mathbf{X} - \sqrt{N}\mathbf{U}\|^2 \\ &\text{s.t. } \mathbf{U}^*\mathbf{U} = \mathbf{I} \\ &\quad |x(n)| = 1, \ n = 1, \ldots, N. \end{aligned} \qquad (27)$$

As in the section "The CAN Algorithm," a cyclic approach is adopted. X is first initialized by a randomly generated unimodular sequence. The criterion in (27) is then iteratively minimized by fixing X to compute U, then fixing U to compute X (and so on, until a given stop criterion is satisfied). During this iterative process, both U and X have closed-form updating formulae (see [26] for details). Although CA does not follow an FFT-based approach, we can also extend the CAN approach described in the section "The CAN Algorithm" to design sequences whose correlation lags are only minimized over a region of interest. Further details can be found in [24].

SEQUENCE SETS

Many applications, such as MIMO radar and code division multiple access (CDMA) systems, require a set of sequences with both good auto- and cross-correlation properties. We can extend the single sequence scenario, which only considers autocorrelation, to the multiple sequence case as follows.

For a set of M unimodular sequences $\{x_m(n)\}$ ($m = 1, \ldots, M$ and $n = 1, \ldots, N$), the cross-correlation between the kth and sth sequence at time lag l is defined as



[FIG4] Overlaid cross-correlations for a CA sequence set with *M* = 4, *N* = 256, and *P* = 30.

$$r_{ks}(l) = \sum_{n=l+1}^{N} x_k(n)x_s^*(n-l) = r_{sk}^*(-l) \qquad (28)$$

$$k, s = 1, \ldots, M \text{ and } l = 0, \ldots, N-1.$$

The ISL, which now must consider both the auto- and cross-correlations, can be extended to the multiple waveform case as

$$\mathrm{ISL}_{\mathrm{MIMO}} = \sum_{k=1}^{M}\sum_{l=1}^{N-1}|r_{kk}(l)|^2 + \sum_{k=1}^{M}\sum_{\substack{s=1\\ s\neq k}}^{M}\sum_{l=0}^{N-1}|r_{ks}(l)|^2. \qquad (29)$$

Minimization of the ISL in (29) can be performed, for all delays, using an FFT-based approach, which allows for efficient computation and permits the design of longer sequences. This approach parallels the CAN formulation reviewed in the section "The CAN Algorithm," and we refer the reader to [27] for further details. Similarly, when the maximum lag considered is less than the sequence length, the CA approach described in the section "The CA Algorithm" can be directly applied to the multiple sequence case. More information can be found in [27], [32], and [26].

In Figure 4, we provide the cross-correlations for a set of $M = 4$ CA sequences with length $N = 256$. We consider $P = 30$ correlation lags (we are only interested in minimizing $\{r_{ks}(l)\}$ in (28) for $|l| < 30$); we overlay the set of seven cross-correlations (i.e., between the first and second waveform and the first and third waveform). As evidenced, the cross-correlations are well below −250 dB in the region of interest. The autocorrelations of the sequences (not shown) have similar, near-zero behavior in the region of interest.

THE PERIODIC CORRELATION CASE

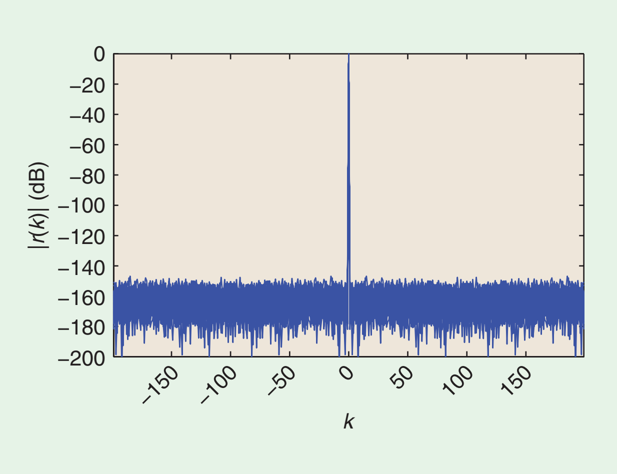
As discussed in the section "A Review of Existing Waveforms," extensive literature exists on the design of signals with good periodic properties. Sequences having low periodic autocorrelations are useful for such applications as CDMA systems [33] and ultrasonic imaging [34]. Further, for applications involving multiple waveforms, sequence sets with good periodic auto and cross-correlations are often desirable. For example, in asynchronous CDMA systems, low periodic autocorrelation improves synchronization and low periodic cross-correlation reduces interference from other users. In this section, we briefly describe the extension of CA to the design of CAZAC sequences of arbitrary length N (referred to as periodic CAN, or PeCAN). Unlike many existing CAZAC waveforms, periodic CA sequences do not have a closed-form expression, which is certainly desirable in many covert applications (for example, covert underwater communications [31]).

We can replace the matrix X in (25) with

$$\mathbf{X} = \begin{bmatrix} x(1) & x(N) & \cdots & x(2) \\ x(2) & x(1) & & x(3) \\ \vdots & \vdots & & \vdots \\ x(N) & x(N-1) & \cdots & x(1) \end{bmatrix}_{N\times N}, \qquad (30)$$

where each column is a shifted version of the sequence $\{x(n)\}_{n=1}^{N}$. The matrix product $\mathbf{X}^*\mathbf{X}$ will now include periodic correlations at all lags, and the minimization problem in (27) (and subsequent cyclic solution) follows as before. For the design of longer sequences, we can instead adopt an FFT-based approach, similar to the one used by CAN in the section "The CAN Algorithm." We refer the readers to [35] and [36] for more details on the cyclic design of sequences with good periodic properties.

We show the superimposed (periodic) autocorrelations of 50 periodic CAN sequences of length $N = 200$ in Figure 5. In our simulation, we actually generated 100 sequences using random initializations; the 50 sequences shown represent those with the lowest ISL. As shown in Figure 5, the sidelobes for each of these sequences is below -140 dB (and can be considered zero in practice).



[FIG5] Overlaid autocorrelations of 50 periodic CAN sequences of length $N = 200$.

RECEIVER DESIGN

In the section "Transmit Waveform Design," we described several different waveforms, all designed to provide a high MF and thus allow for better clutter suppression at the receiver. For some cases, however, even a careful construction of the radar's transmit waveforms, when coupled with a matched filter at the receiver, still might not provide sufficient sidelobe reduction. To address these situations, we now turn our attention to the receiver stage of an active sensing system. We begin our discussion by reviewing the matched filter and by motivating the need for more advanced receiver designs.

MATCHED FILTER

A matched filter is applied, in many applications, to improve the SNR properties of the received signal (see, e.g., [12] and [37]). Ideally, a matched filter works by amplifying the signal of interest component in the received signal and by reducing the signal's noise component, which is usually assumed to be uncorrelated with the transmitted sequence(s). In the presence of stochastic additive white noise, in fact, a matched filter provides the highest SNR performance. If the transmitted waveform, or waveform set, has good correlation properties, a matched filter will also serve to weaken the reflected signals from targets in neighboring range cells to the one of interest.

After the matched filter is applied to $\mathbf{y}_{r'}$, the least-squares estimate for the reflection coefficient $\kappa_{r',l'}$ is then given by

$$\hat{\kappa}_{r',l'} = \frac{\sum_{n=1}^{N} \tilde{x}_{l'}^{*}(n)\, y_{r'}(n)}{\sum_{n=1}^{N} |\tilde{x}_{l'}(n)|^2} = \frac{\tilde{\mathbf{x}}_{l'}^{*}\mathbf{y}_{r'}}{\tilde{\mathbf{x}}_{l'}^{*}\tilde{\mathbf{x}}_{l'}}. \tag{31}$$

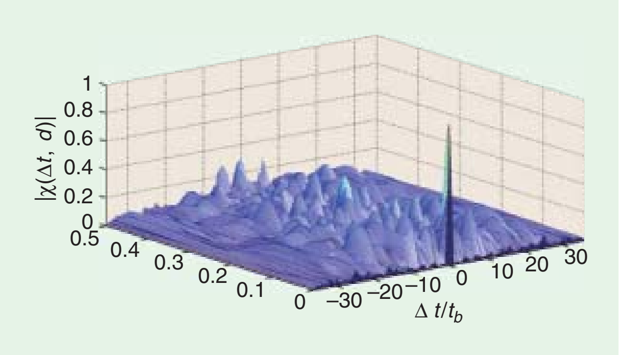
Similar estimates can be generated for the other targets in the scene by reformulating the model for the received signal in (3) (so that $\mathbf{y}_{r'}$ is aligned with the return from a range bin of interest r' for $r' = 1, \ldots, R$).

If there were no interference terms in (3) (i.e., if $\kappa_{r,l} = 0$ for any $\{r, l\} \neq \{r', l'\}$), then the matched filter would provide a highly accurate estimate of $\kappa_{r',l'}$. When interference terms (clutter) are present in the received signal, which is commonly the case in practice, then the performance of the matched filter for estimation will depend directly on the correlation properties of the transmitted sequence(s). In the section "Transmit Waveform Design," we described several cyclic approaches that can be used to design sequences (or sequence sets) with low correlations. When the correlation region of interest is small enough, as exemplified in Figure 4, we are often able to synthesize sequences (or sequence sets) with nearly zero sidelobes over the region of interest. When we wish to minimize the correlation across all sequence lags, or when Doppler effects are nonnegligible, however, it is not possible to design waveforms (or waveform sets) with zero ISL.

The autocorrelation of a waveform $s(t)$ represents the matched filter's temporal response to a target with negligible Doppler shift (a stationary target relative to the radar). If a target is moving, we have to instead consider the ambiguity function of the signal [12]

$$|\chi(\Delta t, d)| = \left| \int_{-\infty}^{\infty} s(t)\, s^{*}(t + \Delta t)\, e^{j2\pi dt}\, dt \right|, \tag{32}$$

where Δt again represents the relative time delay and d represents the Doppler shift of a target. Unlike the autocorrelation, the volume (sidelobes) underneath the ambiguity function for any sequence is constrained to unity (when we normalize by the energy in the signal). In Figure 6, we show the three-dimensional representation of the ambiguity function of a CAN signal (initialized with a random sequence) with length $N = 36$ (where $t_b = \tau/N$ denotes the length of each subpulse). As we can see, the ambiguity function of the CAN waveform resembles a "thumbtack" in shape. Although a "thumbtack" form is desirable, since this shape can lead to improved Doppler resolution, the total volume underneath the function remains fixed. Since we



[FIG6] **Ambiguity function of a length $N = 36$ CAN function.**

are unable to design a sequence (or a set of sequences) that has zero sidelobes for all time delays and Doppler shifts in (32), we instead seek to replace the matched filter with more advanced receiver designs.

IV RECEIVE FILTER

The instrumental variables (IV) method (also called a mismatched filter), a more general approach for estimating $\kappa_{r',l'}$, can be used to significantly lower sidelobes at the cost of a reduced SNR [38], [39], [40]. Temporarily neglecting Doppler effects (so that $L = 1$, $\omega_1 = 0$, and $\tilde{\mathbf{x}}_1 = \mathbf{x}$), the IV estimate of $\kappa_{r'}$ is given by

$$\hat{\kappa}_{r'} = \frac{\mathbf{z}^*\mathbf{y}_{r'}}{\mathbf{z}^*\mathbf{x}}, \tag{33}$$

where $\mathbf{z}$ denotes the IV receive filter. In the case $\mathbf{z} = \mathbf{x}$, then (33) reduces to the matched filter estimate of $\kappa_{r'}$. In general, and unlike the matched filter, the elements of $\mathbf{z}$ are not restricted to be unimodular, since this vector is only designed for the purposes of estimation. Also, we note that IV filters can be precomputed offline. From a computational standpoint, therefore, IV certainly offers minimal burden to the receiver, as the complexity of its application is comparable to that of the matched filter. We assume herein that $\mathbf{z}$ is a vector of length N, although, by padding the transmit waveform with zeros, a longer IV vector could be designed to improve sidelobe reduction even more (at a cost of further reduced SNR).

We consider the IV formulation given in [40]. The goal of the IV approach is to find a signal $\mathbf{z}$ that minimizes the ISL, which, in the negligible Doppler case, is given by

$$\text{ISL}_{\text{IV}} = \frac{\sum_{k=-(N-1),\,k\neq 0}^{N-1} |\mathbf{z}^*\mathbf{J}_k\mathbf{x}|^2}{|\mathbf{z}^*\mathbf{x}|^2}. \tag{34}$$

By applying the Cauchy-Schwartz inequality, the minimum value of ISL_{IV} was shown to be achieved when $\mathbf{z} = \mathbf{R}_{\text{IV}}^{-1}\mathbf{x}$, where

$$\mathbf{R}_{\text{IV}} = \sum_{k=-N+1,\,k\neq 0}^{N-1} \mathbf{J}_k\mathbf{x}\mathbf{x}^*\mathbf{J}_k^T. \tag{35}$$

In this way, an IV receive vector, in the absence of Doppler effects, can be designed to reduce sidelobes to near zero levels. When motion is present in the scene, however, an IV filter can fail to provide satisfactory results.

We will now assume that the Doppler shifts of the targets in the scene $\{\omega_l\}_{l=1}^{L}$ are assumed to lie within an uncertainty interval denoted by $\Omega = [\omega_a, \omega_b]$ (where $\omega_b > \omega_a$ and where we choose L such that $\{\omega_l\}_{l=1}^{L}$ covers Ω). Since no knowledge is assumed of the targets' Doppler shifts, other than that they belong to Ω, the ISL criterion in (34) is rewritten as [40]

$$\text{ISL}_{\text{IV, D}} = \sum_{\substack{k=-(N-1)\\ k\neq 0}}^{N-1} \left(\frac{1}{\omega_b - \omega_a}\right) \frac{\int_{\Omega} |\mathbf{z}_{l'}^*\mathbf{J}_k\tilde{\mathbf{x}}(\omega)|^2 d\omega}{|\mathbf{z}_{l'}^*\tilde{\mathbf{x}}_{l'}|^2}, \tag{36}$$

where $\mathbf{z}_{l'}$ refers to the receive filter for Doppler bin l' and $\tilde{\mathbf{x}}(\omega)$ denotes the Doppler shifted waveform (according to Doppler frequency ω). When the Doppler uncertainty interval Ω becomes larger, the minimum achievable value of $\text{ISL}_{\text{IV, D}}$ could become significantly greater than that of ISL_{IV}. Intuitively, this is due to the fact that the designs based on $\text{ISL}_{\text{IV, D}}$ are more conservative, as they try to optimize the ISL metric averaged over the entire set Ω. For this reason, the IV approach does not perform well when Doppler effects are nonnegligible.

ITERATIVE ADAPTIVE APPROACH

To provide higher resolution in the nonnegligible Doppler case, at the cost of increased computational complexity at the receiver, we now explore a more advanced estimation technique. The iterative adaptive approach (IAA), first presented in [41], was shown to offer improved resolution and interference rejection performance. IAA is a nonparametric and user parameter-free weighted least-squares algorithm. In [41], IAA was shown to perform well for applications in channel estimation, radar and sonar range-Doppler imaging, and passive array sensing. Whereas some data-adaptive algorithms require a significant number of snapshots to obtain accurate target estimates, IAA was shown to achieve good performance even with a single data vector. We briefly summarize the algorithm here.

Consider the model for $\mathbf{y}_{r'}$ in (3). The goal of IAA is to minimize the following weighted least-squares cost function with respect to a target of interest $\kappa_{r',l'}$

$$\|\mathbf{y}_{r'} - \kappa_{r',l'}\tilde{\mathbf{x}}_{l'}\|^2_{\mathbf{Q}^{-1}_{r',l'}}, \tag{37}$$

where $\|\mathbf{u}\|^2_{\mathbf{Q}^{-1}} \triangleq \mathbf{u}^*\mathbf{Q}^{-1}\mathbf{u}$. The interference covariance matrix for a target of interest $\kappa_{r',l'}$ is denoted by $\mathbf{Q}_{r',l'}$, and is defined

$$\mathbf{Q}_{r',l'} = \mathbf{R}_{\text{IAA}}(r') - |\kappa_{r',l'}|^2\tilde{\mathbf{x}}_{l'}\tilde{\mathbf{x}}_{l'}^*, \tag{38}$$

where

$$\mathbf{R}_{\mathrm{IAA}}(r') = \sum_{r=-N+1}^{N-1} \sum_{l=1}^{L} |\kappa_{r'+r,l}|^2 \mathbf{J}_r \widetilde{\mathbf{x}}_l \widetilde{\mathbf{x}}_l^* \mathbf{J}_r^T. \quad (39)$$

The weighted least-squares estimate for a target of interest $\kappa_{r',l'}$, after some simplification, is given by

$$\hat{\kappa}_{r',l'} = \frac{\widetilde{\mathbf{x}}_{l'}^* (\mathbf{R}_{\mathrm{IAA}}(r'))^{-1} \mathbf{y}_{r'}}{\widetilde{\mathbf{x}}_{l'}^* (\mathbf{R}_{\mathrm{IAA}}(r'))^{-1} \widetilde{\mathbf{x}}_{l'}}, \quad l' = 1, \ldots, L, \quad r' = 1, \ldots, R. \quad (40)$$

Since the estimate in (40) depends on the covariance matrix $\mathbf{R}_{\mathrm{IAA}}(r')$, which in turn depends on the target amplitudes, the algorithm uses an iterative approach, which is summarized in Table 2. The target coefficients are initialized using the matched filter approach outlined in the section "Matched Filter." To estimate targets in other range bins, we simply redefine $\mathbf{y}_{r'}$, which represents the N length signal vector aligned with the received reflection from a range bin of interest r'. IAA typically converges after about ten iterations (which corresponds to $T_{\mathrm{IAA}} = 10$ in Table 2); a local convergence proof for IAA is offered in [42].

ADVANCES IN COMPUTING POWER WILL CONTINUE TO HERALD NEW AND IMPROVED APPROACHES TO WAVEFORM DESIGN.

REGULARIZED IAA

In applications involving multiple receive antennas (which permits the use of steering beams), the angular scanning region, relative to a system's antenna array, might be reduced from the entire range (e.g., −90° to 90°) to a region of interest (e.g., −30° to 30°). Although a reduction in the size of the angular grid would certainly provide computational advantages at the receiver (fewer targets that would require an estimate), such a reduction would inevitably lead to a higher condition number for the covariance matrix $\mathbf{R}$ in (39) (and eventually threaten the invertibility of $\mathbf{R}$). To account for targets that lie outside the scanning region and to also allow for any noise in the received signal, which is not explicitly considered in (39), we might sometimes consider regularization of $\mathbf{R}$ with a diagonal matrix $\boldsymbol{\Sigma}$.

[TABLE 2] IAA FOR RANGE-DOPPLER IMAGING.

```
INITIALIZE ($t = 0$)
$\hat{\kappa}^{(0)}_{r',l'} = \frac{1}{N}\widetilde{x}^*_{l'} y_{r'}, \quad l' = 1, \ldots, L, \quad r' = 1, \ldots, R$
REPEAT ($t = t + 1$)
  FOR $r' = 1, \ldots, R$
    $\mathbf{R}^{(t)}_{IAA}(r') = \sum_{r=-N+1}^{N-1} \sum_{l=1}^{L} |\hat{\kappa}^{(t-1)}_{r'+r,l}|^2 J_r \widetilde{x}_l \widetilde{x}^*_l J_r^T$
    FOR $l' = 1, \ldots, L$
      $\hat{\kappa}^{(t)}_{r',l'} = \frac{\widetilde{\mathbf{x}}^*_{l'}(\mathbf{R}^{(t)}_{\mathrm{IAA}}(r'))^{-1}\mathbf{y}_{r'}}{\widetilde{\mathbf{x}}^*_{l'}(\mathbf{R}^{(t)}_{\mathrm{IAA}}(r'))^{-1}\widetilde{\mathbf{x}}_{l'}}$
    END FOR
  END FOR
UNTIL ($t = T_{\mathrm{IAA}}$)
```

An approach described as IAA-Regularized (IAA-R) was presented in [42] to automatically compute the noise powers in $\boldsymbol{\Sigma}$. In this way, IAA-R fits entirely within the user parameter-free framework of IAA. At a cost of increased computational complexity (since now the noise powers must also be computed iteratively), IAA-R was shown to outperform the original IAA for applications in MIMO radar imaging. Further details and examples can be found in [42].

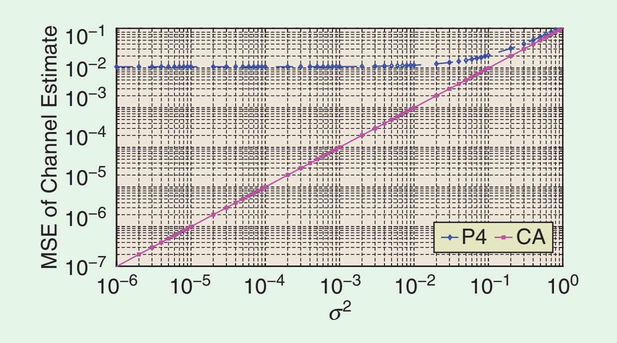
NUMERICAL EXAMPLES

In this section, we will provide several numerical examples to objectively demonstrate the performance of the aforementioned approaches to transmit waveform synthesis in various active sensing applications. Further, we will seek to clarify the advantages and disadvantages of the different receive filters described in the section "Receiver Design."

EXAMPLE 1

First, we aim to illustrate the improved estimation performance that can result by using signals (specifically the CA sequences depicted in the section "The CA Algorithm") with low autocorrelation. Consider an FIR channel impulse response with 40 randomly generated channel taps. Similar to other active sensing applications, the goal of channel estimation is to successfully determine the unknown channel taps. At the transmitter, we adopt a probing pulse of length $N = 200$. The noise in the received signal is assumed to be independent and identically distributed (i.i.d.) complex Gaussian noise with zero mean and variance given by σ^2. Using a matched filter at the receiver to estimate the channel taps, we can compare the performance of a P4 and CA transmit sequence. We assume that the length of the channel is known, so that $P = 40$ in the signal design stage.

In Figure 7, we show the mean-squared error (MSE) of the channel estimate when the noise variance σ^2 is varied from 10^{-6} to 1. We perform 500 Monte Carlo trials for each noise level. Owing to its better autocorrelation properties, the CA



[FIG7] MSE of estimate for FIR channel with $N = 200$.

APPENDIX A

The CAN and PeCAN approaches, which were described in the sections "The CAN Algorithm" and "The Periodic Correlation Case," respectively, are closely related to the GSA that was introduced more than 35 years ago for applications in optics research (note that the cyclic approach described as GSA can also be found in an earlier paper [47], in which a proof of convergence is also provided). In this appendix, we will highlight the similarities and clarify the relationship between these waveform design algorithms and GSA.

GSA

Let $\mathbf{x}$ be an $N \times 1$ vector and consider minimizing the following criterion with respect to $\mathbf{x}$:

$$C(\mathbf{x}) = \sum_{k=1}^{K} [|\mathbf{a}_k^* \mathbf{x}| - d_k]^2, \tag{41}$$

where $d_k \in \mathbb{R}^+$ and $\mathbf{a}_k \in \mathbb{C}^{N \times 1}$ are given and K is an integer that typically satisfies $K \geq N$. In some applications, the vector $\mathbf{x}$ is free to vary in $\mathbb{C}^{N \times 1}$ (see, e.g., [48]). In other applications $\mathbf{x}$ is constrained to a certain subset of $\mathbb{C}^{N \times 1}$, such as to the set of vectors with unimodular elements. To take this fact into account, we let $\mathbf{x} \in S \subseteq \mathbb{C}^{N \times 1}$.

The GSA was introduced in [28] for tackling recovery problems typically involving a sequence and its Fourier transform. When used for problems that can be formulated as in (41), GSA has the following form:

Step 0: Given initial values $\{\psi_k^0\}_{k=1}^K$ ($\{\psi_k\}$ are auxiliary variables; see below for details), iterate Steps 1 and 2 below, for $i = 0, 1, \ldots$ until convergence.

$$\text{Step 1: } \mathbf{x}^i = \arg\min_{\mathbf{x} \in S} \sum_{k=1}^{K} |\mathbf{a}_k^* \mathbf{x} - d_k e^{j\psi_k^i}|^2. \tag{42}$$

Step 2: $\psi_k^{i+1} = \arg(\mathbf{a}_k^* \mathbf{x}^i)$ and $i \leftarrow i + 1$.

The algorithm is useful when the minimization problem in Step 1 has a closed-form solution, which is obviously true for $S = \mathbb{C}^{N \times 1}$, but also for some significant instances of constraint sets (see, e.g., [24] and [35]).

Note that [28] proposed the above algorithm on heuristic grounds, without any reference to the minimization of $C(\mathbf{x})$ in (41). However, it was later shown in [49] that GSA is a minimization algorithm for (41) that has the appealing property of monotonically decreasing the criterion as the iteration proceeds. A simple proof of this fact is as follows:

$$\begin{aligned} C(\mathbf{x}^i) &= \sum_{k=1}^{K} [\,|\mathbf{a}_k^* \mathbf{x}^i| - d_k]^2 = \sum_{k=1}^{K} |\mathbf{a}_k^* \mathbf{x}^i - d_k e^{j\psi_k{}^{i+1}}|^2 \\ &\geq \sum_{k=1}^{K} |\mathbf{a}_k^* \mathbf{x}^{i+1} - d_k e^{j\psi_k{}^{i+1}}|^2 \\ &\geq \sum_{k=1}^{K} |\mathbf{a}_k^* \mathbf{x}^{i+1} - d_k e^{j\psi_k{}^{i+2}}|^2 = C(\mathbf{x}^{i+1}), \end{aligned} \tag{43}$$

where the first inequality is due to Step 1 and the second inequality is due to Step 2 (these inequalities are strict if the solutions computed in Steps 1 and 2 are unique, which is usually the case in applications).

The calculation in (43) provides a way to motivate GSA as a minimization algorithm for $C(\mathbf{x})$. In the following, we outline a way to derive GSA as a minimizing procedure for $C(\mathbf{x})$.

Let $\boldsymbol{\psi}$ denote a $K \times 1$ vector of auxiliary variables and let $D(\mathbf{x}, \boldsymbol{\psi})$ be a function which has the property that

$$\min_{\boldsymbol{\psi}} D(\mathbf{x}, \boldsymbol{\psi}) = C(\mathbf{x}). \tag{44}$$

Then, under rather general conditions, the $\mathbf{x}$ that minimizes $C(\mathbf{x})$ is the same as the $\mathbf{x}$ obtained from the minimization of $D(\mathbf{x}, \boldsymbol{\psi})$ with respect to both $\mathbf{x}$ and $\boldsymbol{\psi}$. Evidently, for this approach to be useful the minimization of $D(\mathbf{x}, \boldsymbol{\psi})$ should be easier to handle than that of $C(\mathbf{x})$. To use the above idea in the present case of (41), we let

$$D(\mathbf{x}, \boldsymbol{\psi}) = \sum_{k=1}^{K} |\mathbf{a}_k^* \mathbf{x} - d_k e^{j\psi_k}|^2, \tag{45}$$

where $\boldsymbol{\psi}$ is the vector made from $\{\psi_k\}_{k=1}^K$. We note that the above function has the required property

transmit sequence significantly outperforms the P4 signal for smaller values of σ^2.

EXAMPLE 2

When the number of range bins in an application exceeds the length of the transmit sequence, the CAN approach (described in the section "The CAN Algorithm") can be used to generate waveforms with minimum correlation values across all lags. For this example, we will use a SISO radar system to perform range profiling of a scene. In doing so, we seek to highlight the CAN waveforms, as well as to motivate the need for better receiver design. We consider a scenario with $R = 512$ equally spaced range bins. We place three stationary (negligible Doppler effects) targets in the scene: one target at range bin 200 with amplitude -7 dB, one target at range bin 308 with amplitude -17 dB, and one target at range bin 320 with amplitude 0 dB. The transmit waveforms are designed with $N = 256$. We will assume circularly symmetric i.i.d. additive complex Gaussian noise with zero-mean and variance σ^2. The SNR, in decibels, is defined as $\text{SNR} = 10\log_{10}(1/\sigma^2)$, and is set to 20 dB. True target locations are indicated on each of the figures using an "O."

The result using a Frank sequence and a matched filter at the receiver is shown in Figure 8(a). As evidenced, the two stronger targets are successfully identified using this scheme. The third, weaker target, however, appears within the sidelobes of the strongest target, and the matched filter does not produce a peak at the true target location. In Figure 8(b), we again use a matched filter, but now transmit a CAN waveform. Since $R > N$ for this imaging example, we choose a CAN sequence, as opposed to a CA sequence, to effectively minimize all correlation lags in the waveform synthesis stage. For this case, sidelobes are reduced, and a peak is now discernible at the location of the weakest target. We use CAN waveforms for the remaining figures.

We adopt an IV receive filter (with length N) in Figure 8(c). Compared to the matched filter result in Figure 8(b),

$$\min_{\boldsymbol{\psi}} D(\mathbf{x},\boldsymbol{\psi}) = \min_{\boldsymbol{\psi}}\sum_{k=1}^{K}[|\mathbf{a}_k^*\mathbf{x}|^2 + d_k^2 - 2|\mathbf{a}_k^*\mathbf{x}|d_k\cos(\arg(\mathbf{a}_k^*\mathbf{x}) - \psi_k)]$$
$$= \sum_{k=1}^{K}[|\mathbf{a}_k^*\mathbf{x}| - d_k]^2 = C(\mathbf{x}). \qquad (46)$$

The minimization of $D(\mathbf{x}, \boldsymbol{\psi})$ with respect to $\mathbf{x}$ (unconstrained as in [48] or constrained as in [24]) for fixed $\boldsymbol{\psi}$ and, respectively, with respect to $\boldsymbol{\psi}$ for fixed $\mathbf{x}$ has simple closed-form solutions. Consequently $D(\mathbf{x}, \boldsymbol{\psi})$, and hence $C(\mathbf{x})$, can be minimized conveniently via a cyclic algorithm in which $\boldsymbol{\psi}$ is fixed to its most recent value and $D(\mathbf{x}, \boldsymbol{\psi})$ is minimized with respect to $\mathbf{x}$, and vice versa. The so-obtained algorithm is nothing but the GSA in (42) and its property in (43) follows immediately from (44) and the fact that the cyclic minimization of $D(\mathbf{x}, \boldsymbol{\psi})$ yields the following monotonically decreasing sequence of criterion values: $C(\mathbf{x}^i) = D(\mathbf{x}^i, \boldsymbol{\psi}^{i+1}) \geq D(\mathbf{x}^{i+1}, \boldsymbol{\psi}^{i+2}) = C(\mathbf{x}^{i+1})$.

The general approach based on (45) can be applied to other problems for which it can lead to algorithms that have little, if anything, in common with GSA (see, e.g., [50]).

CAN AND PeCAN

The central problem dealt with in [24] and [35], as well as in the sections "The CAN Algorithm" and "The Periodic Correlation Case," was the design of a code sequence with impulse-like aperiodic and, respectively, periodic correlations. A main result proved in these papers was the fact that the said problem can be reduced to that of minimizing a criterion of the form

$$\tilde{C}(\mathbf{x}) = \sum_{k=1}^{K}[|\mathbf{a}_k^*\mathbf{x}|^2 - d_k^2]^2, \qquad (47)$$

for a certain K, $\{\mathbf{a}_k\}$, and $\{d_k\}$. In the context of the section "The CAN Algorithm" [see (17)], $K = 2N$, $\mathbf{a}_k = [e^{j\theta_k} \;\; \ldots \; e^{j\theta_k N}]^T$, and $d_k = \sqrt{N}$ (where $\theta_k = 2\pi k/2N$, for $k = 1, \ldots, 2N$).

The criterion in (47) might seem rather similar to $C(\mathbf{x})$ in (41), but in fact there are important differences between these two criteria. A first difference is that (44)–(46) obviously do not hold for $\tilde{C}(\mathbf{x})$. Consequently one cannot derive a GS-type algorithm for (47) by following the approach based on (44) and (45). Of course, we could use a $\tilde{D}(\mathbf{x}, \boldsymbol{\psi})$ defined as

$$\tilde{D}(\mathbf{x}, \boldsymbol{\psi}) = \sum_{k=1}^{K}|(\mathbf{a}_k^*\mathbf{x})^2 - d_k^2 e^{j\psi_k}|^2 \qquad (48)$$

for which it holds that $\min_{\boldsymbol{\psi}} \tilde{D}(\mathbf{x}, \boldsymbol{\psi}) = \tilde{C}(\mathbf{x})$, as required. However, the minimization of $\tilde{D}(\mathbf{x}, \boldsymbol{\psi})$ is not easier than that of $\tilde{C}(\mathbf{x})$.

To get around the above problem, a principal observation made in [24] and [35] was that, under certain conditions, the minimization of (47) is almost equivalent (in a sense specified in [35]) to that of $D(\mathbf{x}, \boldsymbol{\psi})$ in (45). Using this observation and the minimization approach outlined in the paragraph following (46), the CAN and PeCAN algorithms were introduced in [24] [35] for minimizing $D(\mathbf{x}, \boldsymbol{\psi})$ (see the sections "The CAN Algorithm" and "The Periodic Correlation Case," respectively). These algorithms have the same form as the GSA in (42). However, note that now the minimization of $D(\mathbf{x}, \boldsymbol{\psi})$ does not necessarily provide a solution to the problem of minimizing $\tilde{C}(\mathbf{x})$. In particular, a second difference between the criteria $C(\mathbf{x})$ and $\tilde{C}(\mathbf{x})$ is that the algorithms do not guarantee that the criterion $\tilde{C}(\mathbf{x})$ monotonically decreases as the iteration proceeds (only $D(\mathbf{x}, \boldsymbol{\psi})$ is monotonically decreased by each iteration).

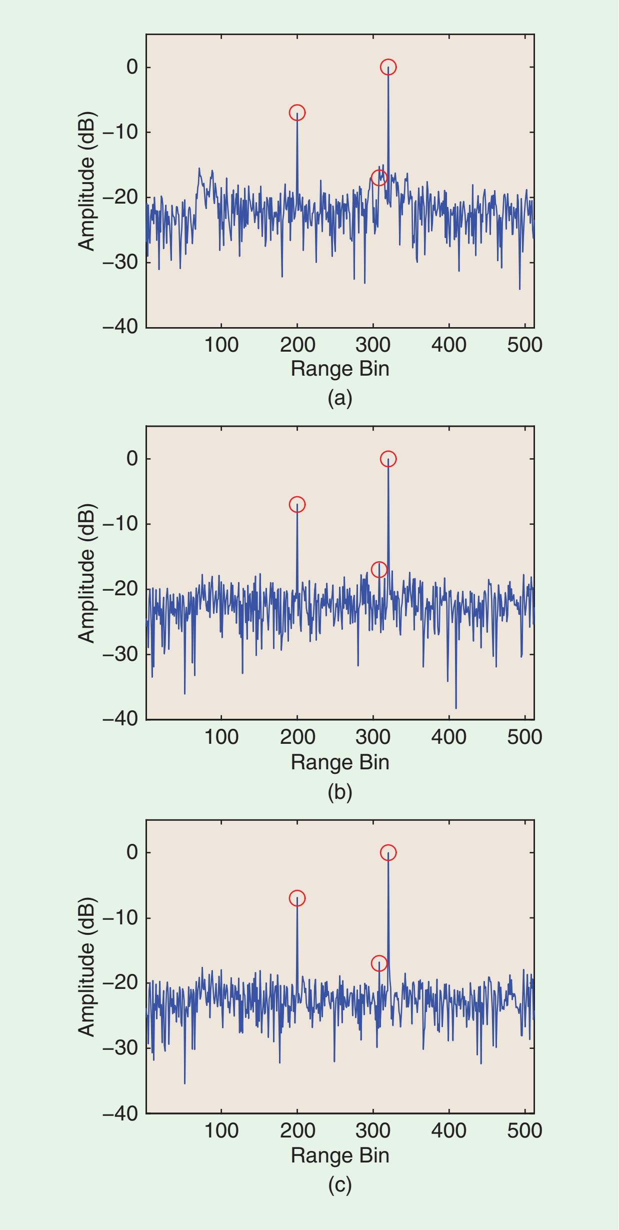
Finally, we remark on the fact that the weighted CAN and multivariate CAN algorithms (introduced in [24] and [27] and reviewed in the sections "The CA Algorithm" and "Sequence Sets," respectively), although related to GSA in their basic principles, have a weaker connection to GSA than CAN and PeCAN. These algorithms, which have been obtained by means of the "almost equivalent" minimization approach mentioned in the previous paragraph, can be viewed as extensions of GSA to problems that have more involved forms than (41) (these problems, as considered in [24] and [27], are associated with the design of sequences with more complex correlations than an impulse-like shaped one).

IV is able to further suppress sidelobe levels and to form well-separated peaks at each of the true target locations (with a negligible increase in computation). IAA (whose result is not shown) achieves similar performance to the IV filter, but at the cost of increased computational efforts at the receiver.

EXAMPLE 3

We now simulate an angle-range synthetic aperture radar (SAR) imaging example using a MIMO antenna scheme (e.g., an airborne radar that scans a ground scene with stationary targets). For this application, we will demonstrate the performance of the IAA approach for receiver design and furthermore showcase the CA sequence sets described in the section "Sequence Sets." To extend the received signal model given in (3) (as well as the receiver designs offered in the section "Receiver Design") to the MIMO case, please refer to [42] and [26]. The MIMO system under consideration contains a uniform linear array with five transmit antennas spaced at $2.5\lambda_0$ and five receive antennas spaced at $0.5\lambda_0$, where λ_0 denotes the carrier wavelength of the system. In this way, i.e., with a sparse transmit array and filled receive array, we effectively create a filled virtual array with $NM = 25$ antennas [43]–[45]. The radar collects data at five positions, with $12.5\lambda_0$ separation between each position. The ground truth consists of 16 targets, with amplitude 0 dB, placed randomly (both in angle and range) within $R = 24$ range bins. True target locations are again indicated on each of the figures using an "O," where now each symbol is colored according to its corresponding amplitude. We let the angular scanning region range from -30° to 30° with 1° grid size. We will use a CA sequence set for transmission with a length of 128. The SNR is 40 dB, where we assume i.i.d. circularly symmetric complex Gaussian noise.

We show the angle-range imaging results in Figure 9. At the receiver, we use a matched filter in Figure 9(a). To improve

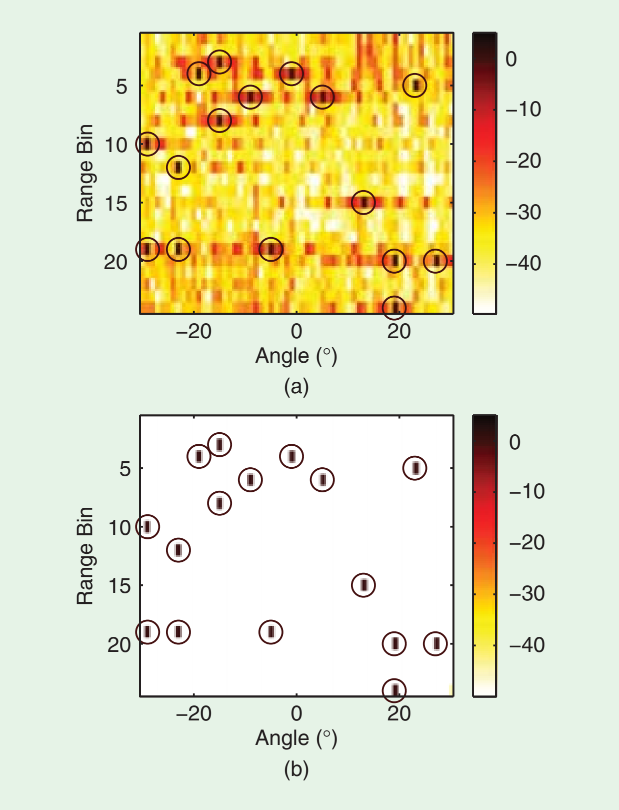


[FIG8] Range profiles for N = 256, SNR = 20 dB, and using (a) a Frank sequence with a matched filter at the receiver, (b) a CAN sequence with a matched filter at the receiver, and (c) a CAN sequence with an IV receive filter. "O" denotes a true target location.

resolution, we apply IAA-R to the received signal in Figure 9(b). Since the scanned angular region is reduced from 180° to only cover a region of interest (for computational purposes), we apply the regularized version of IAA to account for interferences outside the scanning region. As shown, the targets are clearly identifiable using IAA-R (in fact, a perfect result is obtained).

EXAMPLE 4

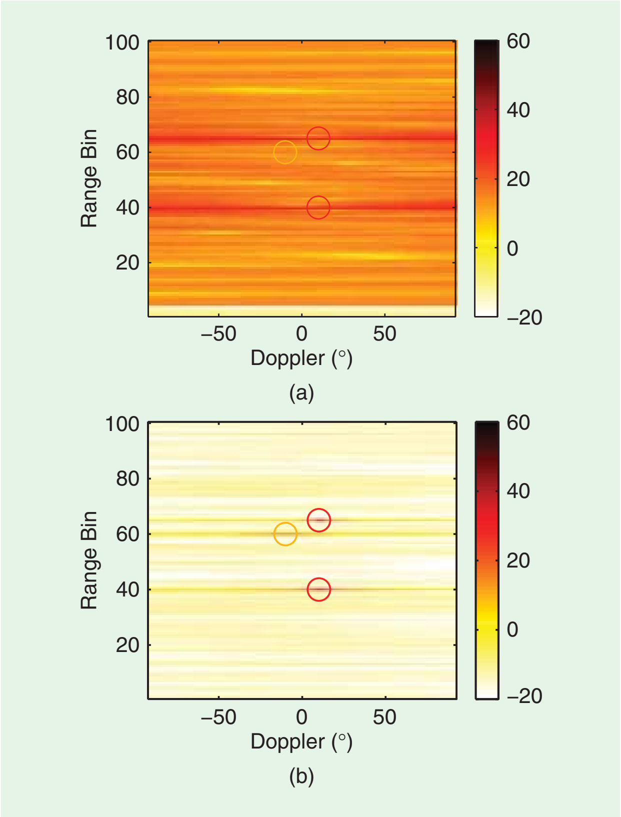
Finally, we will consider range-Doppler imaging using a SISO radar system (or, equivalently for our simulation, a SISO sonar system). By incorporating an example with nonnegligible



[FIG9] MIMO angle-range images for CA transmit sequences with N = 128, SNR = 40 dB, and using (a) a matched filter and (b) IAA-R. Results shown are in dB. True target locations are indicated by "O."

Doppler effects, we hope to demonstrate the need for the IAA algorithm addressed in the section "Iterative Adaptive Approach" (an advanced, more computationally demanding approach to receiver design). The intrapulse Doppler shift of a target is represented by $\Phi_l = \omega_l N(180^\circ/\pi)$, where $l = 1, \ldots, L$. The scene contains $R = 100$ equally spaced range bins and $L = 37$ Doppler bins with 5° separation between bins (we define Ω by setting $\Phi_1 = -90^\circ$ and $\Phi_L = 90^\circ$). We consider three targets in the scene. The first target is located at range bin 60 with Doppler shift -10° and amplitude 10 dB. The second and third targets have Doppler shift 10° and amplitude 30 dB, and are located at range bin 40 and range bin 65, respectively. The SNR is set at 10 dB (relative to a target of amplitude 0 dB and again, assuming circularly symmetric i.i.d. noise). We use a CAN transmit waveform of length $N = 36$.

The imaging result obtained using a matched filter at the receiver is shown in Figure 10(a). As shown, the matched filter fails to provide a peak at the location of the weakest target. The IV filter (whose result is omitted) shows similar performance compared to the matched filter for this nonnegligible Doppler case. The IAA result is shown in Figure 10(b). Compared to the matched and IV filters, IAA significantly reduces sidelobes and produces a peak at each of the true target locations (again, at the cost of increased computation).



[FIG10] Range-Doppler images for a CAN transmit sequence with length $N = 36$, SNR = 10 dB, and using (a) a matched filter and (b) IAA. Results shown are in decibels.

CONCLUSIONS

For decades, researchers have sought to improve the performance of active sensing systems by designing transmission sequences with better correlation properties. Many sequences exhibit perfect periodic correlation, which is desired for some applications in communications and imaging. Other applications, including radar and sonar, demand waveforms with improved aperiodic properties. Due to the difficult computational nature of this problem, the design of sequences with good aperiodic correlation has remained an unsolved and largely evolving research field. In this article, we have provided a brief tutorial of several cyclic algorithms that can be used to efficiently generate sequences and sequence sets with superior auto- and cross-correlations. We described how this cyclic approach to waveform design can be extended to design perfect periodic waveforms. When further improvements in resolution and interference suppression are needed and cannot be met in the signal design stage, better signal processing at the receiver is required. At the expense of a loss in SNR, IV receive filters, which can be precomputed offline, can provide improved performance in the negligible Doppler (stationary target) case compared to a matched filter. When motion is present in the scene, IAA, at the cost of increased computational burden, was shown to produce higher resolution and more accurate target estimates.

Advances in computing power will continue to herald new and improved approaches to waveform design. While we have focused herein on the construction of signals with good correlation properties, this method of sequence design does not account for the Doppler properties of waveforms, which are instead represented by the well-known ambiguity function. The design of signals with specific range-Doppler characteristics is a computationally challenging and application-specific task, and will certainly remain the frontier of research in this field for years to come. In addition, with the increasing popularity of MIMO communications and MIMO radar, we can certainly expect future work to continue to focus on the design of sequence sets with good auto- and cross-correlations, a previously prohibitive (computationally) research area.

ACKNOWLEDGMENTS

This work was supported in part by the SMART Fellowship Program, the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-07-1-0450, the Swedish Research Council (VR), the European Research Council (ERC), the National Science Foundation under grants CCF-0634786, ECCS-0729727, and ECS-0621879; and the Office of Naval Research under grant 00014-09-1-0211.

AUTHORS

William Roberts (WRoberts83@hotmail.com) received his B.Sc. and M.Sc. degrees in electrical and computer engineering from the University of Florida in 2006 and 2007, respectively. He is currently finishing a Ph.D. degree in the Department of Electrical and Computer Engineering at the University of Florida. His research focus includes radar signal processing and spectral estimation.

Hao He (haohe@ufl.edu) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, in 2007, and the M.Sc. degree from the University of Florida, Gainesville, in 2009, both in electrical engineering. He is currently pursuing a Ph.D. degree with the Department of Electrical and Computer Engineering at the University of Florida. His research interests are in the areas of radar/sonar waveform design and spectral estimation.

Jian Li (li@dsp.ufl.edu) received the M.Sc. and Ph.D. degrees in electrical engineering from The Ohio State University, Columbus, in 1987 and 1991, respectively. She is a professor with the Department of Electrical and Computer Engineering, University of Florida, Gainesville. Her current research interests include spectral estimation, statistical and array signal processing, and their applications. She is a Fellow of the IEEE and a Fellow of IET. She is a member of the Sensor Array and Multichannel Technical Committee of the IEEE Signal Processing Society. She is also a member of the editorial board of *IEEE Signal Processing Magazine* and *Digital Signal Processing*.

Petre Stoica (ps@it.uu.se) received his M.Sc. and Ph.D. degrees in automatic control from the Polytechnic Institute of Bucharest in 1972 and 1979, respectively, and an honorary degree in science from Uppsala University in 1993. He is a professor of systems modeling in the Department of Information Technology of Uppsala University in Sweden. His research interests are biomedical signal processing, radar signal processing, wireless communications, echo cancellation, array signal processing, time series analysis, spectral analysis, and system identification. According to the Institute of Science Information, he is one of the 250 most highly cited engineering researchers in the world. He is a Fellow of IEEE. He is also a member of the Royal Swedish Academy of Engineering Sciences, an honorary member of the Romanian Academy, a Fellow of the Royal Statistical Society, and a member of the European Academy of Sciences.

REFERENCES

[1] W. H. Mow, *Sequence Design for Spread Spectrum*. Hong Kong, China: The Chinese Univ. Press, 1995.

[2] T. Helleseth and P. V. Kumar, "Sequences with low correlation," in *Handbook of Coding Theory*, V. S. Pless and W. C. Huffman, Eds. Amsterdam, The Netherlands: Elsevier, 1998, ch. 21, pp. 1765–1853.

[3] N. Levanon and E. Mozeson, *Radar Signals*. Hoboken, NJ: Wiley, 2004.

[4] S. W. Golomb and G. Gong, *Signal Design for Good Correlation: For Wireless Communication, Cryptography, and Radar*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[5] Special Issue on Waveform Agility in Radar Systems, *IEEE Signal Processing Mag.*, vol. 26, no. 1, Jan. 2009.

[6] J. J. Benedetto, I. Konstantinidis, and M. Rangaswamy, "Phase-coded waveforms and their design: The role of the ambiguity function," *IEEE Signal Processing Mag.*, vol. 26, pp. 22–31, Jan. 2009.

[7] R. Calderbank, S. D. Howard, and W. Moran, "Waveform diversity in radar signal processing," *IEEE Signal Processing Mag.*, vol. 26, pp. 32–41, Jan. 2009.

[8] J. Li and P. Stoica, "MIMO radar with colocated antennas: Review of some recent work," *IEEE Signal Processing Mag.*, vol. 24, pp. 106–114, Sept. 2007.

[9] A. H. Haimovich, R. S. Blum, and L. J. Cimini, "Mimo radar with widely separated antennas," *IEEE Signal Processing Mag.*, vol. 25, pp. 116–129, Jan. 2008.

[10] J. Li and P. Stoica, Eds. *MIMO Radar Signal Processing*. Hoboken, NJ: Wiley, 2009.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U,K.: Cambridge Univ. Press, 2005.

[12] M. I. Skolnik, *Radar Handbook*, 2nd ed. New York: McGraw-Hill, 1990.

[13] J. Jedwab, "A survey of the merit factor problem for binary sequences," in *Sequences and Their Applications—SETA 2004* (Lecture Notes in Computer Science, vol. 3486), T. Helleseth, D. Sarwate, H. Y. Song, and K. Yang, Eds. Heidelberg: Springer-Verlag, 2005, pp. 30–55.

[14] T. Høholdt, "The merit factor problem for binary sequences," in *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes* (Lecture Notes in Computer Science, vol. 3857), M. Fossorier, H. Imai, S. Lin, and A. Poli, Eds. Heidelberg: Springer-Verlag, 2006, pp. 51–59.

[15] R. H. Barker, "Group synchronizing of binary digital systems," in *Communication Theory*, W. Jackson, Ed. London, U.K.: Butterworths, 1953, pp. 273–287.

[16] R. Turyn, "On Barker codes of even length," *Proc. IEEE*, vol. 51, no. 9, p. 1256, 1963.

[17] S. Eliahou and M. Kervaire, "Barker sequences and difference sets," *L'Enseignement Math.*, vol. 38, no. 2, pp. 345–382, 1992.

[18] S. W. Golomb, *Shift Register Sequences*. San Francisco, CA: Holden-Day, Inc., 1967.

[19] R. Frank, "Polyphase codes with good nonperiodic correlation properties," *IEEE Trans. Inform. Theory*, vol. 9, pp. 43–45, Jan. 1963.

[20] B. Lewis and F. Kretschmer, Jr., "Linear frequency modulation derived polyphase pulse compression codes," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-18, pp. 637–641, Jan. 1991.

[21] M. Friese and H. Zottmann, "Polyphase Barker sequences up to length 31," *Electron. Lett.*, vol. 30, pp. 1930–1931, Nov. 1994.

[22] A. R. Brenner, "Polyphase Barker sequences up to length 45 with small alphabets," *Electron. Lett.*, vol. 34, pp. 1576–1577, Aug. 1998.

[23] P. Borwein and R. Ferguson, "Polyphase sequences with low autocorrelation," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1564–1567, Apr. 2005.

[24] P. Stoica, H. He, and J. Li, "New algorithms for designing unimodular sequences with good correlation properties," *IEEE Trans. Signal Processing*, vol. 57, pp. 1415–1425, Apr. 2009.

[25] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Englewood Cliffs, NJ: Prentice-Hall, 2005.

[26] J. Li, P. Stoica, and X. Zheng, "Signal synthesis and receiver design for MIMO radar imaging," *IEEE Trans. Signal Processing*, vol. 56, pp. 3959–3968, Aug. 2008.

[27] H. He, P. Stoica, and J. Li, "Designing unimodular sequence sets with good correlations—Including an application to MIMO radar," *IEEE Trans. Signal Processing*, vol. 57, pp. 4391–4405, Nov. 2009.

[28] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.

[29] F. Kretschmer, Jr. and K. Gerlach, "Low sidelobe radar waveforms derived from orthogonal matrices," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 27, pp. 92–102, Jan. 1991.

[30] H. A. Khan, Y. Zhang, C. Ji, C. J. Stevens, D. J. Edwards, and D. O'Brien, "Optimizing polyphase sequences for orthogonal netted radar," *IEEE Signal Processing Lett.*, vol. 13, pp. 589–592, Oct. 2006.

[31] J. Ling, Y. Yardibi, X. Su, H. He, and J. Li, "Enhanced channel estimation and symbol detection for high speed MIMO underwater acoustic communications," *J. Acoust. Soc. Amer.*, vol. 125, pp. 3067–3078, May 2009.

[32] P. Stoica, J. Li, and X. Zhu, "Waveform synthesis for diversity-based transmit beampattern design," *IEEE Trans. Signal Processing*, vol. 56, pp. 2593–2598, June 2008.

[33] N. Suehiro, "A signal design without co-channel interference for approximately synchronized CDMA systems," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 837–841, June 1994.

[34] V. Diaz, J. Urena, M. Mazo, J. Garcia, E. Bueno, and A. Hernandez, "Using Golay complementary sequences for multi-mode ultrasonic operation," in *Proc. IEEE 7th Int. Conf. Emerging Technologies and Factory Automation*, UPC Barcelona, Catalonia, Spain, Oct. 1999, pp. 599–604.

[35] P. Stoica, H. He, and J. Li, "On designing sequences with impulse-like periodic correlation," *IEEE Signal Processing Lett.*, vol. 16, pp. 703–706, Aug. 2009.

[36] H. He, D. Vu, P. Stoica, and J. Li, "Construction of unimodular sequence sets for periodic correlations," in *Proc. 2009 Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1–4, 2009.

[37] G. L. Turin, "An introduction to matched filters," *IRE Trans. Inform. Theory*, vol. 6, pp. 311–329, June 1960.

[38] M. H. Ackroyd and F. Ghani, "Optimum mismatched filters for sidelobe suppression," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 9, pp. 214–218, Mar. 1973.

[39] S. Zoraster, "Minimum peak range sidelobe filters for binary phase-coded waveforms," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 16, pp. 112–115, Jan. 1980.

[40] P. Stoica, J. Li, and M. Xue, "Transmit codes and receive filters for radar," *IEEE Signal Processing Mag.*, vol. 25, pp. 94–109, Nov. 2008.

[41] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer. "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 425–443, Jan. 2010.

[42] W. Roberts, P. Stoica, J. Li, T. Yardibi, and F. Sadjadi, "Iterative adaptive approaches to MIMO radar imaging," *IEEE J. Select. Topics Signal Processing* (Special Issue on MIMO Radar and Its Applications), vol. 4, no. 1, pp. 5–20, Feb. 2010.

[43] K. Forsythe, D. Bliss, and G. Fawcett, "Multiple-input multiple-output (MIMO) radar: Performance issues," in *Proc. 38th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, vol. 1, pp. 310–315, Nov. 2004.

[44] D. W. Bliss and K. W. Forsythe, "Multiple-input multiple-output (MIMO) radar and imaging: Degrees of freedom and resolution," in *Proc. 37th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, vol. 1, pp. 54–59, Nov. 2003.

[45] J. Li, P. Stoica, L. Xu, and W. Roberts, "On parameter identifiability of MIMO radar," *IEEE Signal Processing Lett.*, vol. 14, pp. 968–971, Dec. 2007.

[46] W. Roberts, H. He, X. Tan, M. Xue, D. Vu, J. Li, and P. Stoica, "Probing waveform synthesis and receive filter design for active sensing systems," in *Proc. SPIE Defense, Security and Sensing Conf.*, Orlando, FL, Apr. 2009, pp. 73350H.1–73350H.13.

[47] S. M. Sussman, "Least-square synthesis of radar ambiguity functions," *IRE Trans. Inform. Theory*, vol. 8, pp. 246–254, Apr. 1962.

[48] A. Weiss and J. Picard, "Maximum-likelihood position estimation of network nodes using range measurements," *IET Signal Processing*, vol. 2, no. 4, pp. 394–404, 2008.

[49] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, pp. 2758–2769, Feb. 1982.

[50] H. He, P. Stoica, and J. Li. (2010, June). "Waveform design with stopband and correlation constraints for cognitive radar," in *Proc. 2nd Int. Workshop Cognitive Information Processing*. Elba Island, Italy [Online]. Available: http://plaza.ufl.edu/haohe/papers/SCAN.pdf

[SP]

Daniel Rueckert and Paul Aljabar

Nonrigid Registration of Medical Images: Theory, Methods, and Applications

Medical image registration [1] plays an increasingly important role in many clinical applications, including the detection and diagnosis of diseases, planning of therapy, guidance of interventions, and the follow-up and monitoring of patients. The primary goal of image registration is to find corresponding anatomical or functional locations in two or more images. This has many applications: registration can be applied to images from the same subject acquired by different imaging modalities (multimodal image registration) or at different time points (serial image registration). Both cases are examples of intrasubject registration since the images are acquired from the same subject. Another application area for image registration is intersubject registration, where the aim is to align images acquired from different subjects, e.g., to study the anatomical variability within or across populations.

While rigid registration has become a widely used tool in clinical practice, nonrigid registration has not yet achieved the same level of clinical acceptance. Much recent progress has been made, however, in developing improved nonrigid registration techniques. In this article, we will illustrate some of the advances that have been made over recent decades. We will discuss some of the theoretical aspects of nonrigid registration and describe methods for their implementation. Finally, we will illustrate how common problems in medical imaging, such as motion correction and image segmentation, can be solved using image registration.

Digital Object Identifier 10.1109/MSP.2010.936850

METHODS

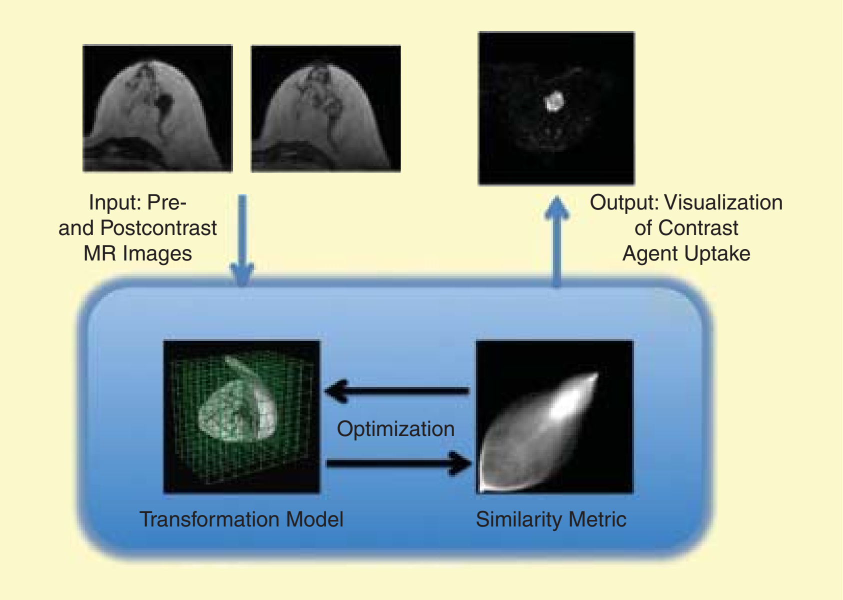
In general, the process of image registration involves finding the optimal geometric transformation that maximizes the correspondences across the images. This involves the following components (see Figure 1):

- *A transformation model* that defines a geometric transformation between the images. There are several classes of nonrigid transformations including parametric and nonparametric models. Some of these models are well suited for small deformations while others can represent large deformations.
- *A similarity metric* that measures the degree of alignment between the images. In cases where features such as landmarks, edges, or surfaces are available, the distances between corresponding features can be used to measure the alignment. In other cases, the image intensities can be directly used to measure the alignment.
- *An optimization method* that maximizes the similarity metric. Like many other problems in medical imaging, nonrigid registration can be formulated as an optimization problem whose goal it is to maximize an associated objective function.

In addition, a careful validation must be performed to assess measures of performance, such as accuracy and robustness, as well as in application-specific terms, such as clinical utility. In the following, we will describe the individual components of nonrigid registration techniques in more detail.

TRANSFORMATION MODELS

The transformation model used in the registration defines how the coordinates



[FIG1] Illustration of the components of a generic nonrigid image registration algorithm: The transformation model, the similarity metric and the optimization technique. This example shows the application of nonrigid registration for the motion correction of contrast-enhanced MR images [2].

1053-5888/10/$26.00©2010IEEE

of two images are related. For a pair of images $\mathcal{I}_A$ and $\mathcal{I}_B$, this is often expressed as a single coordinate transformation $\mathbf{T}$, mapping each point $\mathbf{x}$ in $\mathcal{I}_A$ to an anatomically corresponding location $\mathbf{T}(\mathbf{x})$ in $\mathcal{I}_B$. Some transformations require only translation, rotation, or scaling and, in this case, the output coordinates of $\mathbf{T}(\mathbf{x})$ can be written as a linear combination the input coordinates for some fixed global set of linear weights. In the case of nonrigid registration, no such global linear model can be formulated, and it is common to optimize a spatially varying displacement field $\mathbf{u}$ to express the transformation, i.e., $\mathbf{T}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$. Typical requirements of a nonrigid transformation are that it is smooth and invertible, i.e., that it does not lead to effects such as tearing or collapsing regions to a point. Such requirements reflect the variations in anatomy where changes in size and shape are common but changes in topology are rare.

As an illustration of some of the aspects of nonrigid transformation models, we describe the widely used free-form deformations (FFDs) that were developed within the computer graphics and computer-aided design communities and now also have an established role in medical image registration [2]. An FFD is defined by a set of displacement vectors associated with the points of a discrete three-dimensional (3-D) lattice. A blend of the vectors is used to define the displacement at a general location in the image with nearer vectors having a greater influence. The blending weights are determined by a weighting function and spline functions, such as B-splines, are often used.

FFDs are an example of parametric transformations and contrast with nonparametric transformations where a displacement vector is associated with every voxel in the image (a voxel is a volume element, the 3-D analogy of a pixel). FFDs are geometric in their construction, but it is possible to derive transformations from more physical models such as fluid [3], diffusion [4], or elastic models [5]. FFDs are also an example of small deformation models that are suitable for modeling, say, gradual changes in anatomy. In some applications, such as the deformations of cardiac muscle, a large deformation model, such as that derived from a flow field, can be more appropriate.

In the case of fluid-based transformation models, the registration no longer seeks to optimize the displacements at each location directly but instead estimates a velocity field that is used to provide the displacement. In this case, the corresponding points $\mathbf{x}$ and $\mathbf{T}(\mathbf{x})$ represent the start and end points of a flow determined by the velocity. This can be a challenge to optimize, especially given that the velocity field may be allowed to vary over time as well as space.

SIMILARITY METRICS

The second component of a registration algorithm is the registration basis that measures the degree of alignment of the images. The two main approaches are feature-based and voxel-based similarity measures. Feature-based registration approaches usually utilize points, lines, or surfaces and aim to minimize the distance between the corresponding features in the images. An advantage of feature-based registration is that it can be used for both mono- and multimodality registration, but the need for a feature extraction step, in the form of landmark detection or segmentation, can be onerous. Moreover, any error during the feature extraction stage, whether manual or automated, will adversely affect the registration and cannot be recovered at a later stage. It is possible to avoid such errors by using the image intensities directly without the need for feature extraction. This relies on voxel-based similarity measures that aim to measure the degree of shared information in the image intensities. This is relatively simple in the case of mono-modality registration but more complex for multimodality registration. Over the last decade, voxel-similarity measures have become the method of choice for measuring image alignment, largely due to their robustness and accuracy.

The simplest statistical measure of image similarity is based on the squared sum of intensity differences (SSD) between images $\mathcal{I}_A$ and $\mathcal{I}_B$,

$$\mathcal{S}_{SSD} = -\frac{1}{n}\sum(\mathcal{I}_A(\mathbf{x}) - \mathcal{I}_B(\mathbf{T}(\mathbf{x})))^2, \qquad (1)$$

where $\mathbf{x}$ is a point in image $\mathcal{I}_A$, $\mathbf{T}(\mathbf{x})$ is the corresponding location in $\mathcal{I}_b$ and n is the number of voxels in the overlap region. This measure is based on the assumption that both imaging modalities have the same characteristics. If the images are correctly aligned, the difference between them should be zero except for noise, and the SSD measure can be shown to be optimal if this noise is Gaussian. Since this similarity measure assumes that the imaging modalities are identical, their application is restricted to mono-modal applications.

The assumption of identical imaging modalities can, however, be too restrictive. A more general approach assumes a linear relationship between the image intensities. In this case, the similarity between both images can be expressed by the normalized cross correlation (NCC) shown in (2) at the bottom of the page, where μ_A and μ_B correspond to the average voxel intensities in each image. While more flexible than SSD, the application of this similarity measure is nevertheless largely restricted to mono-modal registration tasks.

There has been significant interest in measures of alignment based on the information content or entropy of the registered images. An important

$$\mathcal{S}_{CC} = \frac{\sum(\mathcal{I}_A(\mathbf{x}) - \mu_A)(\mathcal{I}_B(\mathbf{T}(\mathbf{x})) - \mu_B)}{\sqrt{\left(\sum \mathcal{I}_A(\mathbf{x}) - \mu_A\right)^2 \left(\sum \mathcal{I}_B(\mathbf{T}(\mathbf{x})) - \mu_B\right)^2}}, \qquad (2)$$

component of these methods is the feature space of the image intensities that may be interpreted as a joint probability distribution. A simple way of visualizing this feature space is by accumulating a two-dimensional histogram of the co-occurrences of intensities in the two images for each trial alignment (Figure 2). By varying the degree to which the images are aligned, it can be shown that the feature space disperses as misalignment increases and that each image pair has a distinctive feature space signature at alignment.

In an information theoretic framework, the information content of images $\mathcal{I}_A$ and $\mathcal{I}_B$ can be defined by their Shannon-Wiener entropy

$$H(\mathcal{I}_A) = -\sum_a p(a) \log p(a) \qquad (3)$$

and

$$H(\mathcal{I}_B) = -\sum_b p(b) \log p(b), \qquad (4)$$

where $p(a)$ is the probability that a voxel in image $\mathcal{I}_A$ has intensity a and $p(b)$ is the probability that a voxel in image $\mathcal{I}_B$ has intensity b. The joint entropy $H(\mathcal{I}_A, \mathcal{I}_B)$ of the overlapping region of images $\mathcal{I}_A$ and $\mathcal{I}_B$ may be defined by

$$H(\mathcal{I}_B, \mathcal{I}_B) = -\sum_a \sum_b p(a,b) \log p(a,b), \qquad (5)$$

where $p(a,b)$ is the joint probability that a voxel in the overlapping region of image $\mathcal{I}_A$ and $\mathcal{I}_B$ has values a and b, respectively.

To quantify image alignment, one can use measures from information theory such as mutual information (MI) [6], [7]. MI is defined in term of entropies as

$$\mathcal{S}_{MI}(\mathcal{I}_A;\mathcal{I}_B) = H(\mathcal{I}_A) + H(\mathcal{I}_B) - H(\mathcal{I}_A,\mathcal{I}_B) \qquad (6)$$

and should be maximal at alignment. MI is a measure of how one image "explains" the other but makes no assumption of the functional form or relationship between image intensities in the two images. Studholme [8] showed that MI can be affected by the degree of overlap between two images. Studholme [8] and Maes et al. [6] suggested the use of normalized MI (NMI) as an alternative measure one form of which may be written

$$\mathcal{S}_{NMI}(\mathcal{I}_A;\mathcal{I}_B) = \frac{H(\mathcal{I}_A) + H(\mathcal{I}_B)}{H(\mathcal{I}_A,\mathcal{I}_B)}. \qquad (7)$$

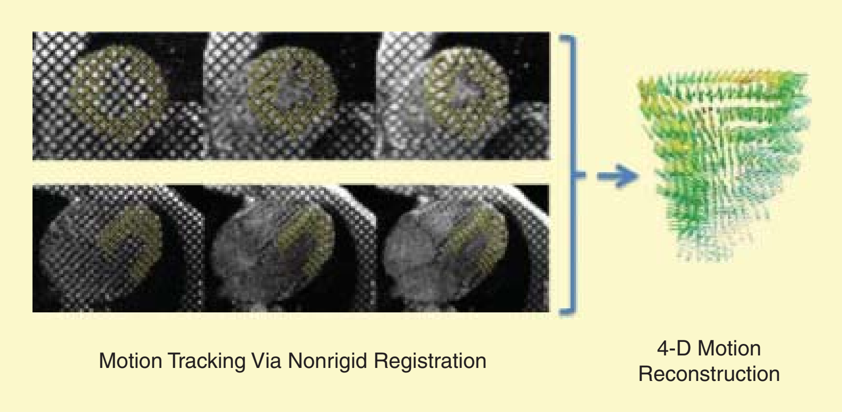
IMPLEMENTATION: OPTIMIZATION AND INTERPOLATION

The registration task seeks to identify the transformation parameters that maximize the similarity measure derived from the two images. In certain special cases, such as the rigid registration of pairs of corresponding landmarks, it is possible to analytically estimate the optimal transformation (in a least squares sense). Such an example is, however, exceptional as the majority of registrations are voxel-based registrations and these typically rely on numerical methods to find the optimal parameters.

THE PRIMARY GOAL OF IMAGE REGISTRATION IS TO FIND CORRESPONDING ANATOMICAL OR FUNCTIONAL LOCATIONS IN TWO OR MORE IMAGES.

In the case of nonrigid transformations, the number of parameters or degrees of freedom can be very high. For example, an FFD with a cubic control point lattice and ten control points along each side has 10 × 10 × 10 × 3 = 3,000 parameters to optimize with respect to the similarity metric. Many optimization methods, such as Newton or conjugate descent approaches, require the estimation of the similarity measure's gradient with respect to the parameters and second-order methods may also require an estimate of its Hessian. For some similarity measures, such as $\mathcal{S}_{SSD}$ or $\mathcal{S}_{CC}$, an explicit expression for the gradient may be derived. This can remain possible for more complex entropy-based measures such as $\mathcal{S}_{MI}$, but the computational overhead of evaluating them can make numerical schemes of gradient approximation more attractive. In the case of the Hessian, even numerical estimation can be computationally expensive for every iteration and techniques for avoiding its direct estimation are often exploited.

As a function of the transformation parameters, the similarity metric defines a hypersurface in a typically high-dimensional space (3,000 dimensions in the earlier example). This presents a challenge to the optimization as the surface is likely to be highly nonconvex with multiple local maxima and a number of approaches have been developed to help identify a globally optimal set of parameters. Notable examples include coarse-to-fine approaches, where image pyramids at different scales are used instead of the original images with transformations at coarser levels of detail optimized first and used as an initialization for the subsequent stages with finer details. As an alternative to gradient-based


Motion Tracking Via Nonrigid Registration
4-D Motion Reconstruction

[FIG2] Tracking change over time in a sequence of tagged MR images for the 4-D estimation of myocardial motion.

optimization methods, it is also possible to discretize the parameter space entirely and apply techniques such as linear programming.

One aspect of many registration approaches is that they are expressed asymmetrically. To evaluate the similarity measure for the images, it is common to loop over the voxels in the first ("target") image, identifying the corresponding locations in the second ("'source") image under the current transformation estimate. Each target-source intensity pair is then used to estimate the similarity. It is unlikely for each location in the target voxel lattice to correspond directly to a source voxel and it is much more likely to correspond to an intermediate location within the source voxel lattice. This necessitates the interpolation of one of the images (the source), hence the asymmetry in the model.

If, for example, a linear interpolation scheme is applied, the source image effectively undergoes a low-pass filtering step and the associated loss of detail can have an effect on the registration. In practice however, such effects are small, especially if an image pyramid is used in a multiresolution scheme. Other more complex interpolation schemes can of course be applied, such as sinc interpolation or higher-order splines, but there is a tradeoff between interpolation accuracy and computational burden that needs to be taken into account.

Some approaches attempt to symmetrize the registration by looping over both source and target voxels, using the inverse transformation to obtain interpolated values from the target image. Again, computational load becomes an issue here as there may be a significant cost to inverting the transformation, and it may be difficult to accurately estimate its inverse.

VALIDATION OF REGISTRATION

Prior to clinical use, medical image registration algorithms need to be validated in terms of their accuracy in establishing correspondence between images. However, the validation of registration performance usually suffers from the lack of knowledge as to whether, how much, and where patient movement has occurred between and even during scanning procedures, and whether such movement affects the clinical usefulness of the data. To maintain clinical usefulness, and to inherently improve patient treatment and health care, it is therefore vital to ensure that registration is successful.

A registration method can be assessed by independent evaluation in the absence of a ground truth correspondence estimate. An initial visual inspection allows for a qualitative assessment of registration performance, which can be complemented by quantitative checks for robustness and consistency. Robustness checks establish the measurement precision by testing the bias and sensitivity after, for example, adding noise or choosing different starting estimates. Consistency checks assess the capability of a registration technique to find circular transformations based on a registration circuit but can be sensitive to bias and may not be applicable to noninvertible transformations generated by many nonrigid registration methods. Nonetheless, consistency checks have been successfully used for intramodality rigid body registration applications, e.g., for serial magnetic resonance (MR) imaging of the brain. The methods available to an expert observer performing a visual assessment of registration performance include the inspection of subtraction images, contour or segmentation overlays, alternate pixel displays, or viewing anatomical landmarks. These approaches have been applied to rigid registration, and since they involve inspection of the entire volume domain of the image pair, can be extended to nonrigid registration. For nonrigid registration, expert visual assessment is an important step toward clinical acceptance and routine use but locally implausible deformations, not readily picked up by observers, represent a significant challenge. Nonetheless, visual assessment often forms the first and last line of defence of any image registration validation.

[A REGISTRATION METHOD CAN BE ASSESSED BY INDEPENDENT EVALUATION IN THE ABSENCE OF A GROUND TRUTH CORRESPONDENCE ESTIMATE.]

In the absence of a ground truth transformation, registration accuracy can be studied by setting up a gold standard transformation. For example, the retrospective registration evaluation project (RREP) used skull-implanted markers in patients undergoing brain surgery to derive a gold standard transformation for multimodality rigid-body image registration of the head to compare different established rigid registration methods [9]. For nonrigid registration validation, extrinsic markers cannot easily be used as they would need to be implanted into soft tissue. In an alternative approach [10], a biomechanical motion simulator was introduced that modeled physically plausible deformations of soft tissue for clinically realistic motion scenarios in an application to contrast-enhanced MR mammography. This motion simulator was designed to be independent of the image registration and transformation model used.

Finally, the segmentation of anatomical structures provides the means to measure structure overlap or surface distances before and after registration, but cannot provide insight into the registration accuracy away from the structure's boundary, or along its outline. If, however, the objective of the registration is to propagate (and hence automate) segmentation, segmentation quality can be used as a surrogate measurement. For example in [11] a number of nonrigid registration methods were compared for intersubject brain alignment based on their segmentation quality and a number of carefully annotated image databases (e.g., http://www.nirep.org/) are becoming available that can be used to establish the accuracy of nonrigid registration methods on the basis of carefully delineated image structures.

APPLICATIONS OF NONRIGID REGISTRATION

TRACKING CHANGES OVER TIME

It is possible to apply the registration of medical images to identify changes in anatomies over time. The changes studied may last for short time intervals, such as muscle deformations in a cardiac cycle, or they may last for years, such as gradual atrophy in an aging brain. Neural atrophy is an example of change that is diffuse and subtle compared with, say, the rapid and dramatic growth of organs during fetal development. Each of these scenarios present its own challenges when registration is used to characterize the associated change. In a longitudinal approach to measuring change, serially acquired images of a single individual may be registered to identify changes in the anatomy that may be of clinical interest. This approach is in routine use for identifying atrophy of the brain due to aging or disease.

In the case of serial images of the head, rigid registration and subtraction can provide a good indication of changes. Other forms of serial images, such as sequences of a beating heart, require nonrigid registration to characterize the deformation of tissue over time (see Figure 2). In the case of serial brain MR images of elderly subjects, nonrigid registration may still be applied and the geometric properties of the nonrigid transformation between the images can provide quantitative local estimates of the tissue expansion or contraction that has taken place between the scans. This helps to identify which structures in the brain are most susceptible to pathology.

In a cross-sectional approach, images are acquired from a number of subjects at varying temporal stages and intersubject registration may be used to factor out variability across subjects and to subsequently identify the salient changes in the images of the whole cohort over time. Example applications are the generation of a four-dimensional (4-D) spatio-temporal atlas of an aging anatomy.

MORPHOMETRY AND SEGMENTATION

Morphometry can generally be described as the study of shape and, in the context of medical images, it can represent the direct comparison or modeling of the shapes of anatomical structures. D'Arcy Thompson used nonrigid coordinate transforms to compare the forms of biological organisms in his seminal work at the start of the 20th century. The computational power now available and the tools of nonrigid registration allow the comparison of anatomies in general and anatomical structures in particular to be carried out systematically and on a large scale. For example, the expected shape of an anatomical structure and its variation over a population can be estimated and this information can be used in a clinical setting to decide if a particular anatomy is representative or pathological.

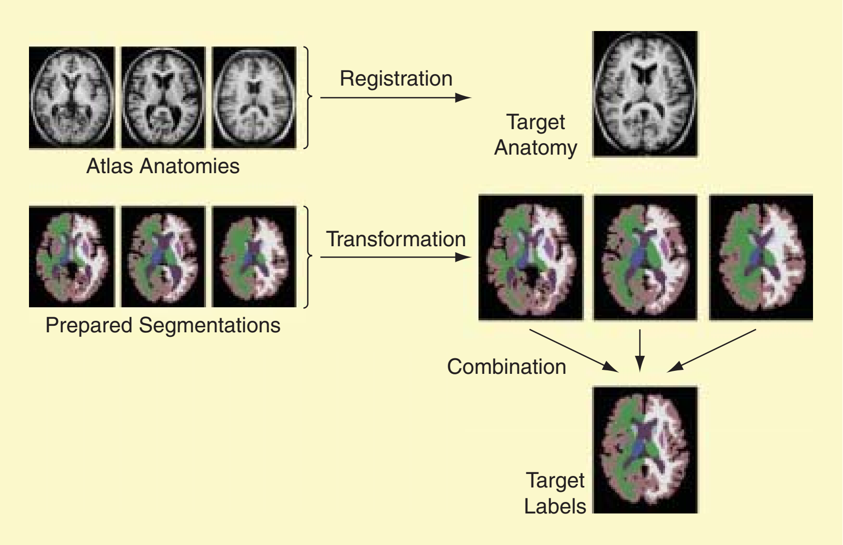
[RECENT ADVANCES IN GPU TECHNOLOGY HAVE THE POTENTIAL TO SIGNIFICANTLY ACCELERATE NONRIGID REGISTRATION AND OFFER THE POSSIBILITY OF NEAR REAL-TIME REGISTRATION.]

Nonrigid registration also enables more indirect approaches to morphometry such as the neuroimaging methods known as voxel-based morphometry (VBM) and deformation-based morphometry (DBM). In VBM, variations in tissue density are identified across a cohort of subjects after aligning all images to the a common template. The registrations used for alignment only aim to correct global changes in the anatomy. After alignment to a common template and subsequent smoothing, variations in tissue densities are used to identify group differences. DBM seeks to characterize the anatomical variation in a population in a similar fashion, i.e., by registering all images to a common template. However, in contrast to VBM, it typically uses a much more detailed registration of the different images. As a result of this, the differences between the anatomies are no longer visible in the images but are instead encoded in the transformations that align them. By studying the geometric properties of the deformations, such as their locally varying Jacobian tensors, group differences can be identified.

Whether studying the shape properties of specific anatomical structures, or applying VBM or DBM, a segmentation step is needed. This can be used to delineate the particular anatomical structure of interest or to identify the tissue or region of interest where the analysis is to be carried out. When delineating anatomical structures, a trained human expert can provide very accurate segmentations although this is a time-consuming and costly exercise. Automated segmentation is, by contrast, much faster and easier to carry out at the expense of a loss of accuracy.

Registration can be used to bridge the gap between the manual and automatic approaches through an approach termed "atlas-based segmentation." In this context, an atlas is represented by an anatomical image together with an expert manual labeling of the structure of interest. When a new image of a different anatomy is obtained, registration can be used to align the atlas anatomy to the new image. The resulting transformation may then be used to transform the atlas label to the new image. The transformed label can then be treated as a segmentation estimate of the new image and, in a sense, the expertise of the human rater has been propagated automatically to the target image.

Atlas-based segmentation is, however, prone to errors that can arise from a variety of sources such as errors in the original labeling or errors in the estimated transformation between atlas and target. It has been shown that a multiatlas approach to segmentation can overcome much of this error and provide very accurate structural segmentations. In this state-of-the-art approach, instead of using single atlas, a repository of atlases are stored and each is separately registered with the target. The resulting set of transformations is used to

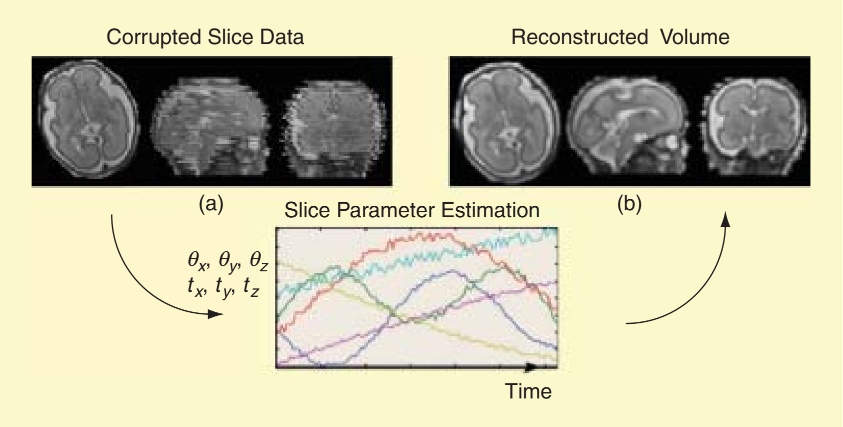


[FIG3] An example of multiatlas segmentation of brain MR images.

propagate the labels for each atlas to the space of the target image. After propagation, a consensus segmentation of the target is obtained by combining the propagated labels using some scheme (see Figure 3). A simple vote scheme in which the label for a voxel (in or out of the structure) is determined according to the majority of the propagated labels has been shown to be highly effective.

MOTION CORRECTION

There are a number of different imaging modalities available to the medical community. MR imaging in particular has a number of attractive properties in that it can distinguish between different types of soft tissue without exposing the subject to ionizing radiation. A drawback of MR imaging is the length of the acquisition time and, when the subject moves, this can lead to artifacts in the acquired image. The acquisition of a MR image represents a tradeoff among various factors such as resolution, the quality of contrast between tissues, and the signal-to-noise ratio (SNR). For example, it is possible to rapidly acquire a volumetric image but the resulting SNR will be low. A high SNR volume requires a long acquisition time in which the subject is more likely to move. It is possible, however, to acquire slice data rapidly and with reasonable quality, although a single slice only provides a restricted view of the anatomy.



[FIG4] Registration for the correction of motion artifacts: (a) successively acquired slices of a moving subject are geometrically inconsistent. (b) Registration is used to estimate the motion parameters of each slice and reconstruct a consistent volume.

A particular and recent application in which images are affected by motion has been the in-utero imaging of fetal subjects and registration may be used to correct for the resulting artifacts. In this approach, a number of parallel slices of the fetus are acquired. Each individual slice is acquired quickly enough for motion to be negligible and represents a high-resolution snapshot through the subject. The fetus is, however, likely to move during the time required to acquire all slices. This means that, while the slices are all parallel relative to the scanner, they are no longer parallel relative to the fetal anatomy. After acquisition, it is possible to use an iterative registration and reconstruction scheme to correct for the mismatch in geometry among all the slices and to estimate a 3-D volumetric reconstruction of the fetal subject's anatomy (see Figure 4).

If the head is the focus of the scan, it can be assumed that all the slices are related to the "true" underlying volume by a rigid alignment. The task is to estimate the transformation parameters for each slice. Once estimated, the relative orientations of the slices are known and it is possible to reconstruct the original signal in three dimensions using a scattered data interpolation approach. In practice, the two main steps (slice parameter estimation and volume reconstruction) are carried out in an iterative and interleaved fashion. It is also possible to adopt a multiresolution coarse-to-fine approach where the early iterations estimate the anatomy at lower spatial frequencies and more detail is recovered as the iterations proceed.

SUMMARY AND OUTLOOK

As we have illustrated in this article, the nonrigid registration of medical images is a versatile tool that is widely used, both in clinical applications (e.g., motion correction and image fusion) as well as a tool for biomedical research (e.g., to study populations or disease progression in clinical trials). In contrast to rigid registration, the development of nonrigid registration techniques is very much an area of ongoing research, and most algorithms are still in the early stages of evaluation and validation. The speed of nonrigid registration algorithms is one drawback of most algorithms, making their clinical use difficult. However, recent advances in GPU technology have the potential to significantly accelerate nonrigid registration and offer the possibility of near real-time registration. However, another drawback is the lack of a generic gold standard for assessing and evaluating the success of nonrigid registration algorithms. Future developments in this area will need to address both issues.

AUTHORS

Daniel Rueckert (D.Rueckert@imperial.ac.uk) is a professor of visual information processing in the Department of Computing, Imperial College London.

Paul Aljabar (Paul.Aljabar@imperial.ac.uk) is a senior research fellow in in the Department of Computing, Imperial College London.

REFERENCES

[1] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, Eds, *Medical Image Registration*. Boca Raton, FL: CRC, 2001.

[2] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Non-rigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, 1999.

[3] M. F. Beg, M. I. Miller, and A. T. L. Younes, "Computing metrics via geodesics on flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[4] J.-P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, 1998.

[5] R. Bajcsy and S. Kovačič, "Multiresolution elastic matching," *Comput. Vis. Graph. Image Process.*, vol. 46, pp. 1–21, Apr. 1989.

[6] F. Maes, A. Collignon, D. Vandermeulen, G. Marechal, and R. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, 1997.

[7] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[8] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1998.

[9] J. B. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, Jr., R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandain, X. Pennec, M. E. Noz, G. Q. Maguire, Jr., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, no. 4, pp. 554–566, 1997.

[10] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of non-rigid image registration using finite element methods: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 238–247, 2003.

[11] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[12] R. Chandrashekara, R. H. Mohiaddin, and D. Rueckert, "Analysis of 3D myocardial motion in tagged MR images using nonrigid image registration," *IEEE Trans. Med. Imag.*, vol. 23, pp. 1245–1250, Oct. 2004.

[SP]

[from the EDITOR]

continued from page 2

whether we should be making changes to embrace them.

Consider first research collaboration. Many of us are involved in long-distance collaborations, taking advantage of inexpensive online communication tools such as Skype. Usually these collaborations start "offline" and our expectation is for online tools to help us maintain them. Maybe we should be looking beyond mere communication and start thinking about tools that will allow us to create, write, and experiment in a distributed and collaborative manner. Perhaps we should even be thinking about how online social networks may help us identify collaborators and define new problems. We are already able to edit papers jointly, and of course open source code is an example of collaborative development. The relevant core tools for shared editing [such as Concurrent Versioning System (CVS)] are well established and widely used. But this could be extended to broader areas: for example, our data sets could be more readily shared and, beyond that, even our experimental settings could be open to all (see, for example, www.myexperiment.org). I look forward to sharing not only the final output of our work (our publications), but also the tools, data sets, and experiments that we include in our results. The recent Netflix competition may be an example of what is becoming possible: teams were able to compare results quickly, with the same data set, they could combine forces to improve their overall score, and the end result might have been better than what a single group working in isolation could have achieved given the same amount of time.

Recommendations are equally important and promising. Consider the publication process. The reviews play a quality control role, but once a paper is published, or even easier once it appears online (say, on Arxiv), then what? So many papers, so little time. Even in new and niche areas, it soon becomes impossible to keep up with *all* the work. And, whether we like to admit it or not, we already rely on recommendations (from colleagues and students, citations by others, and blogs) or on reputation (if an author's previous work was good, we have an incentive to read what they have been up to recently). Why not take this a step further? Let readers rate the papers after they appear, let them comment, critique, and make it easier for everybody to find the really important work. Whether you believe in crowd wisdom (average ratings for a paper) or trust established reviewers (followed on Twitter), there should be ways to improve how we find interesting work and move our research forward.

All these changes may not alter significantly the scientific method or how we convey new knowledge: don't expect many new results to be summarized in a tweet. On the other hand, journals as we know them, with their monthly compilations of unrelated work (that just happened to be accepted around the same time), may well become obsolete sooner than we think. Wherever new results are published, the authors may have to get used to sharing credit with those who help everybody else identify the importance of their work. [SP]

A. Or

[dsp **TIPS&TRICKS**]

Zhi Shen

Improving FIR Filter Coefficient Precision

"DSP Tips and Tricks" introduces practical design and implementation signal processing algorithms that you may wish to incorporate into your designs. We welcome readers to submit their contributions. Contact Associate Editors Rick Lyons (R.Lyons@ieee.org) or C. Britton Rorabaugh (dspboss@aol.com).

There is a method for increasing the precision of fixed-point coefficients used in linear-phase finite impulse response (FIR) filters to achieve improved filter performance. The improved performance is accomplished without increasing either the number of coefficients or coefficient bitwidths. At first thought, such a process does not seem possible, but this article shows exactly how this novel filtering process works.

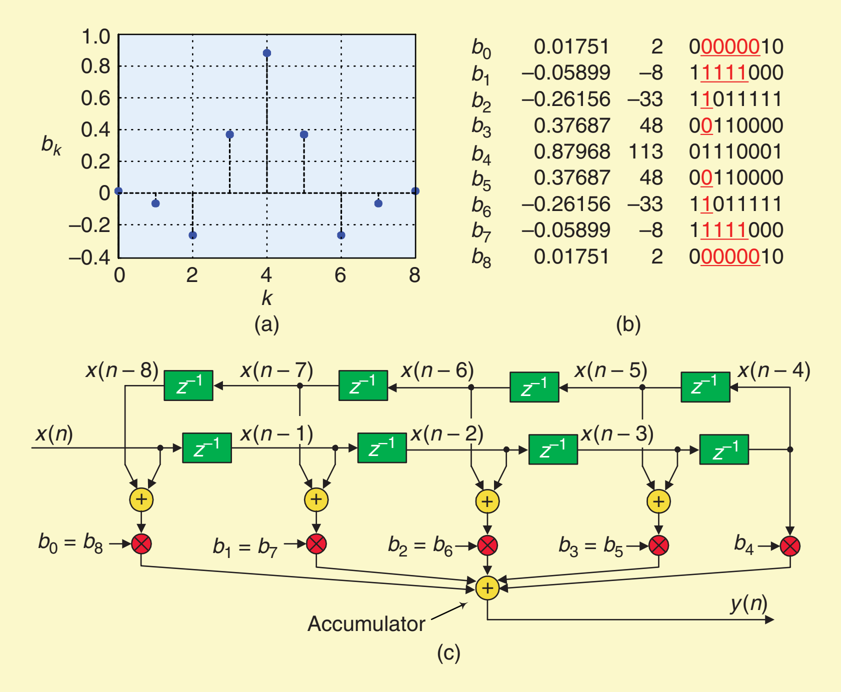
TRADITIONAL FIR FILTERING

To describe our method of increasing FIR filter coefficient precision, let's first recall a few characteristics of traditional linear-phase tapped-delay line FIR filter operation.

Consider an FIR filter whose impulse response is shown in Figure 1(a). For computational efficiency reasons (reduced number of multipliers), we implement such filters using the folded tapped-delay line structure shown in Figure 1(c) [1].

The filter's b_k floating-point coefficients are listed in the second column of Figure 1(b). When quantized to an 8-b two's-complement format, those coefficients are the decimal integers and binary values shown in the third and fourth columns, respectively, in Figure 1(b).

Compared to b_4, the other coefficients are smaller, especially the outer coefficients such as b_0 and b_8. Because of the fixed bitwidth quantization, many high-order bits of the low-amplitude coefficients, the red-font underscored bits in the fourth column of Figure 1(b), are the same as the sign bit. These bits are wasted because they have no effective (no weight) in the calculations. If we can remove those wasted bits (consecutive bits adjacent to, and equal to, the sign bit), and replace them with more significant coefficient bits, we will obtain improved numerical precision for the low-amplitude beginning and ending coefficients.



[FIG1] **Generic linear-phase FIR filter: (a) impulse response, (b) coefficients, and (c) structure.**

Replacing a low-amplitude coefficient's wasted bits with more significant bits is the central point of our FIR filtering trick—of course some filter architecture modification is needed as we shall see. So let's have a look at a generic example of what we call a "serial" implementation of our trick.

SERIAL IMPLEMENTATION

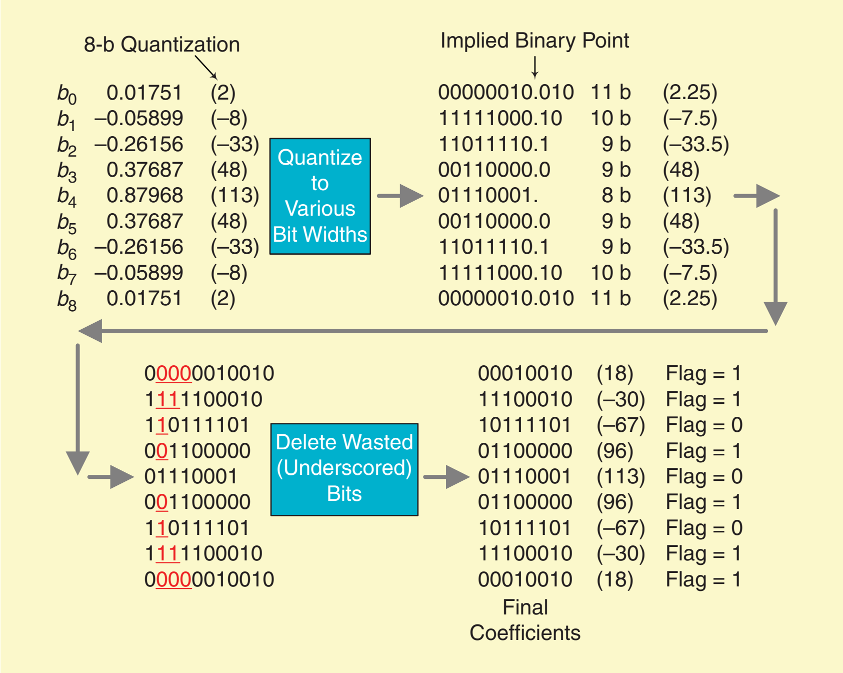
As a simple example of replacing wasted bits, we list the Figure 1(b) b_k coefficients as the floating-point numbers in the upper left side of Figure 2. Assume we quantize the maximum-amplitude coefficient, b_4, to 8 b. In this FIR filter trick we quantize the lower-amplitude coefficients to larger bitwidths than the maximum coefficient (b_4) as shown on

Digital Object Identifier 10.1109/MSP.2010.936777

the upper right side of Figure 2. (The algorithm used to determine those variable bitwidths is discussed later in this article.) Next, we eliminate the appropriate wasted bits, the red-font underscored bits in the lower left side of Figure 2, to arrive at our final 8-b coefficients shown on the lower right side of Figure 2.

Appended to each coefficient is a flag bit that indicates whether that coefficient used one more quantization bit than the previous, next larger, coefficient.

Now, you may say: "Stop! You can't do this. The outer coefficients are left shifted, so they are enlarged, and the product accumulations are changed. Using these modified coefficients, the filter results will be wrong!" Don't worry, we correct the filter results by modifying the way we accumulate products. Let's see how.


8-b Quantization
Implied Binary Point
Quantize to Various Bit Widths
Delete Wasted (Underscored) Bits
Final Coefficients

b_0	0.01751	(2)
b_1	–0.05899	(–8)
b_2	–0.26156	(–33)
b_3	0.37687	(48)
b_4	0.87968	(113)
b_5	0.37687	(48)
b_6	–0.26156	(–33)
b_7	–0.05899	(–8)
b_8	0.01751	(2)

00000010.010	11 b	(2.25)
11111000.10	10 b	(–7.5)
11011110.1	9 b	(–33.5)
00110000.0	9 b	(48)
01110001.	8 b	(113)
00110000.0	9 b	(48)
11011110.1	9 b	(–33.5)
11111000.10	10 b	(–7.5)
00000010.010	11 b	(2.25)

00000010010	00010010	(18)	Flag = 1
1111100010	11100010	(–30)	Flag = 1
110111101	10111101	(–67)	Flag = 0
001100000	01100000	(96)	Flag = 1
01110001	01110001	(113)	Flag = 0
001100000	01100000	(96)	Flag = 1
110111101	10111101	(–67)	Flag = 0
1111100010	11100010	(–30)	Flag = 1
00000010010	00010010	(18)	Flag = 1

[FIG2] Filter coefficients for serial implementation.

The coefficients and flag bits from Figure 2 are used in the serial implementation shown in Figure 3. The data registers in Figure 3 represent the folded delay-line elements in Figure 1(c). This implementation is called "serial" because there is only one multiplier and, when a new $x(n)$ input sample is to be processed, we perform a series of multiplications and accumulations (using multiple clock cycles) to produce a single $y(n)$ filter output sample.

For an N-tap FIR filter, where N is odd, due to our folded delay-line structure only $(N+1)/2$ coefficients are stored in the coefficient read-only memory (ROM). (When N is even, $N/2$ coefficients are stored.) Crucial to this FIR filter trick is that when processing a new $x(n)$ input sample, the largest coefficient, b_4, is applied to the multiplier prior to the first accumulation. Following that is the next smaller coefficient, b_3, and so on. In other words, in this serial implementation the coefficient sequence applied to the multiplier, for each $x(n)$ input sample, is in the order of the largest to the smallest coefficient.

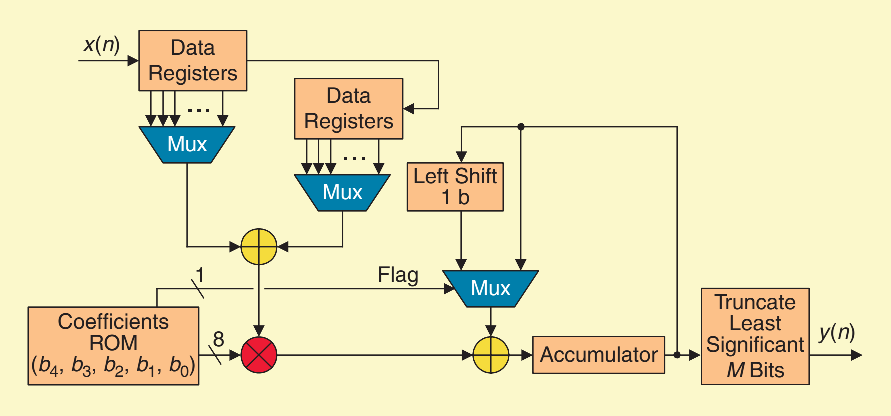
Given these properties, when a new $x(n)$ sample is to be processed we clear the current accumulator value and multiply the sum of the appropriate data registers by the b_4 coefficient. That product is then added to the accumulator. On the next clock cycle we multiply the sum of the appropriate data registers by the b_3 coefficient. If the flag bit of the b_3 coefficient is one, we left shift the current accumulator value and then the current multiplier's output is added to the shifted accumulator value.

THE LEFT SHIFTING OF AN ACCUMULATOR VALUE IS THE KEY TO THIS ENTIRE FIR FILTER TRICK.

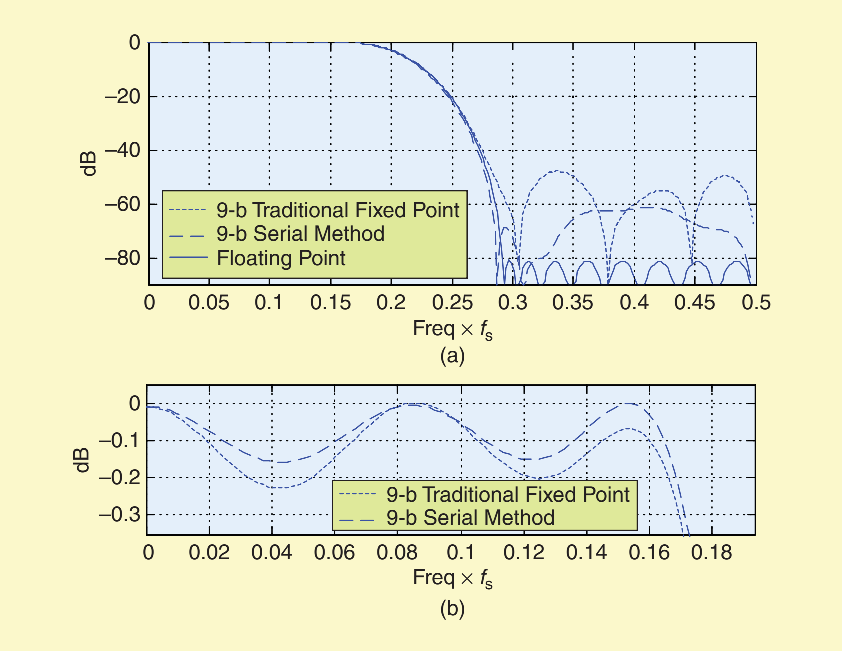
(If the current coefficient's flag bit is zero the accumulator word is not shifted prior to an addition.) We continue these multiplications, possible left shifts, and accumulations for the remaining b_2, b_1, and b_0 coefficients.

The left shifting of an accumulator value is the key to this entire FIR filter trick. To minimize truncation errors due to right shifting a multiplier output word, we preserve precision by left shifting the previous accumulator word.

To maintain our original FIR filter's gain, after the final accumulation we truncate the final accumulator value by discarding its least significant M bits, where M is the total number of flag bits


x(n)
Data Registers
Data Registers
Mux
Mux
Left Shift 1 b
1
Flag
Mux
Coefficients ROM (b_4, b_3, b_2, b_1, b_0)
8
Accumulator
Truncate Least Significant M Bits
y(n)

[FIG3] Serial implementation with 8-b coefficients.



[FIG4] **Low-pass serial method filter frequency responses: (a) full frequency range and (b) passband detail.**

in the ROM memory, to produce a $y(n)$ output sample. Now let's have a look at an actual FIR filter example.

SERIAL METHOD EXAMPLE

Suppose we want to implement a lowpass filter whose cutoff frequency is $0.167f_s$ and whose stopband begins at $0.292f_s$, where f_s is the input data sample rate. If the filter has 29 taps (coefficients), and is implemented with floating-point coefficients, its frequency magnitude response will be that shown by the solid curve in Figure 4(a). Anticipating a hardware implementation using an Altera field-programmable gate array (FPGA) having 9-b multipliers, when using coefficients that have been quantized to 9-b lengths in a traditional manner (with no wasted coefficient bits removed), the filter's frequency magnitude response is the dotted curve in Figure 4(a).

When we use our FIR filter trick's serial implementation, with its enhanced-precision 9-b coefficients (not counting the flag bit) obtained in the manner shown in Figure 2, the filter's frequency magnitude response is the dashed curve in Figure 4(a). We see in the figure that, relative to the traditional fixed point implementation, the serial method provides:

- improved stopband attenuation
- reduced transition region width
- improved passband ripple performance.

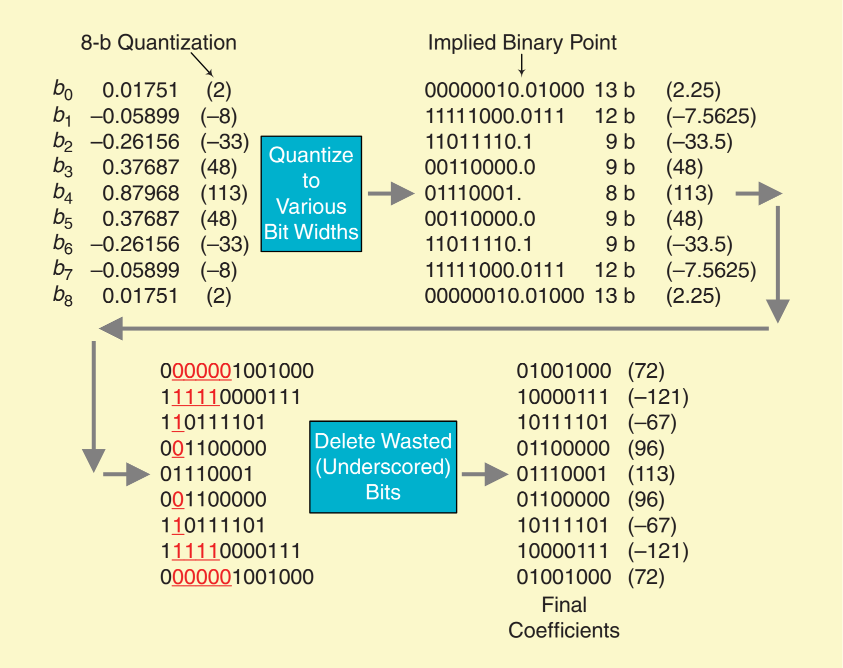
All of these improvements occur without increasing the bitwidths of our filter's multiplier or coefficients, nor the number of coefficients. Because it preserves the impulse response symmetry of the original floating-point filter, the serial implementation filter exhibits phase linearity [2].

It is possible to improve upon the stopband attenuation of our compressed-coefficient serial method FIR filter. We do so by implementing what we call the "parallel method."

PARALLEL IMPLEMENTATION

In the above serial method of filtering, adjacent filter coefficients were quantized to a precision differing by no more that one bit. That's because we use a single flag bit to control the 1-b shifting of the accumulator word prior to a single accumulation. In the parallel method, described now, adjacent coefficients can be quantized to a precision differing by more than 1 b. Figure 5 shows an example of our parallel method's coefficient quantization process.

Again we list the Figure 1(b) b_k coefficients as the floating-point numbers in the upper left side of Figure 5. In this parallel method, however, notice that the expanded quantized b_1 and b_2 words differ by more than one



[FIG5] **Filter coefficients for parallel implementation.**

bit in the upper right side of Figure 5. Coefficients b_2 and b_6 are quantized to 9 b while the b_1 and b_7 coefficients are quantized to 12 b. While we only deleted some of the wasted coefficient bits in Figure 2, in our parallel method all the wasted coefficient bits are deleted. As such, our final 8-b coefficients are those listed in the lower right side of Figure 5.

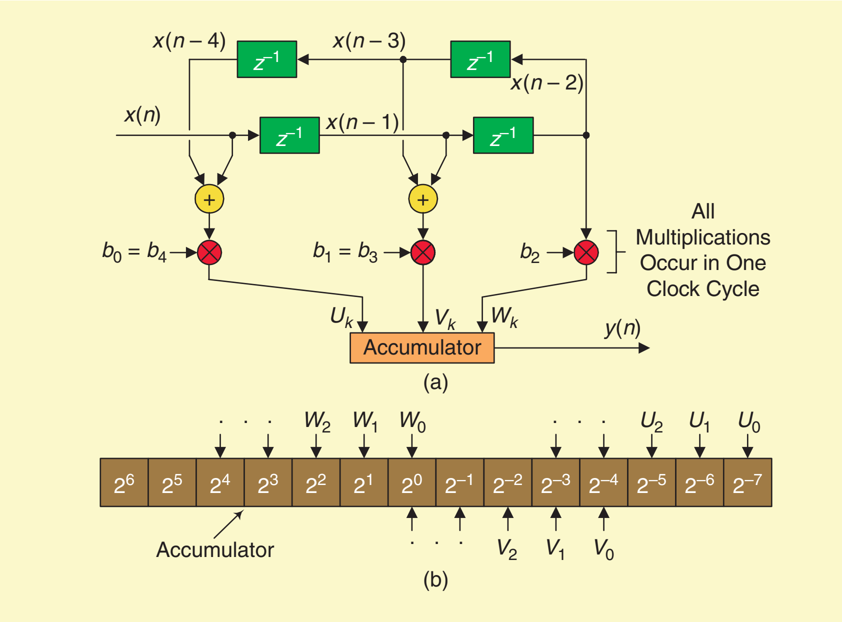
We are all familiar with the operation known as bit extension—the process of extending the bit length of a binary word without changing its value or sign. With that process in mind, we can refer to our trick's operation of removing wasted bits as "bit compression."

Because no flag bits are used in the parallel method, this filtering method is easiest to implement using FPGAs with their flexible multidata bus routing capabilities.

For example, consider the filter structure shown in Figure 6(a) where we perform the three multiplications in parallel (in a single clock cycle) and that is why we use the phrase "parallel method." Instead of shifting the accumulator word to the left as we did in the serial method, here we merely reroute the multiplier outputs to the appropriate bit positions as they are added to the accumulator word as shown in Figure 6(b). In our hypothetical Figure 6 example, if there were four wasted bits deleted from the high-precision b_1 coefficient then the V_k product is shifted to the right by four bits, relative to the W_k product bits, before being added to the accumulator word. If there were seven wasted bits deleted from the high-precision b_0 coefficient, then the U_k product is shifted to the right by 7 b, relative to the W_k product bits, before being added to the accumulator word.

PARALLEL METHOD EXAMPLE

With the solid curve, Figure 7 shows our parallel method's performance in implementing the desired low-pass filter used in the above serial method implementation example. For comparison, we have also included the 9-b traditional fixed point (no bit compression) and the serial method magnitude responses in Figure 7.



[FIG6] **Parallel method implementation: (a) filter structure; (b) accumulator organization.**

The enhanced precisions of the parallel method's quantized coefficients, beyond their serial method precisions, yield improved filter performance. The parallel method of our FIR filter trick achieves a stopband attenuation improvement of 21 dB beyond the traditional fixed-point implementation—again, without increasing the bitwidths of our filter's multipliers or coefficients, nor the number of coefficients.

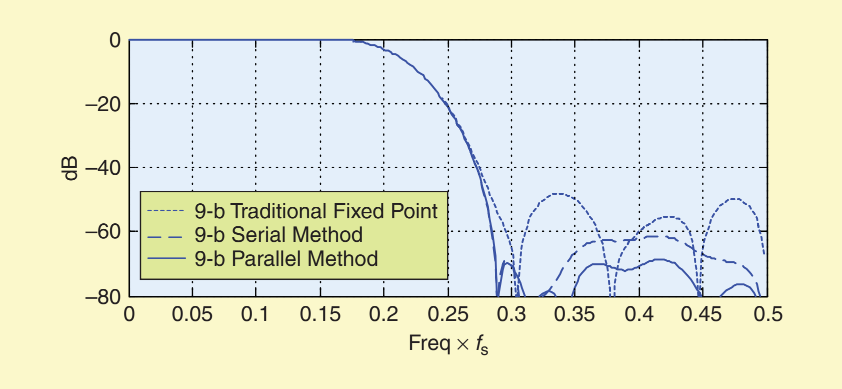
IT IS POSSIBLE TO IMPROVE UPON THE STOPBAND ATTENUATION OF OUR COMPRESSED-COEFFICIENT SERIAL METHOD FIR FILTER.

COMPUTING COEFFICIENT BITWIDTHS

Determining the bitwidths of the quantized filter coefficients in our DSP trick depends on whether you are implementing the serial or the parallel filtering method.

SERIAL METHOD COEFFICIENT QUANTIZATION

In the serial method, let's assume we want our ROM to store coefficients whose bitwidths are integer B (not counting the flag bit).



[FIG7] **Traditional fixed-point, serial method, and parallel method filter frequency responses.**

[TABLE 1] SERIAL METHOD QUANTIZATION EXAMPLE.

COEFFICIENT BEING QUANTIZED	CURRENT SCALE	ALL UNQUANTIZED COEFFICIENTS LESS THAN SCALE/2?	NEW SCALE	K	ROM EQUIVALENT COEFFICIENT BITWIDTH	FLAG BIT	LEFT SHIFT AND ROUND
$b_4 = 0.87968$	1	NO $\|b_4\| > 1/2$	1	7	8	0	$B_4 = \text{ROUND}\ [b_4 \times 2^7] = 113$
$b_3 = 0.37687$	1	YES $\|b_3\|, \|b_2\|, \|b_1\|, \|b_0\| < 1/2$	0.5	8	9	1	$B_3 = \text{ROUND}\ [b_3 \times 2^8] = 96$
$b_2 = -0.26156$	0.5	NO $\|b_2\| > 0.5/2$	0.5	8	9	0	$B_2 = \text{ROUND}\ [b_2 \times 2^8] = -67$
$b_1 = -0.05899$	0.5	YES $\|b_1\|, \|b_0\| < 0.5/2$	0.25	9	10	1	$B_1 = \text{ROUND}\ [b_1 \times 2^9] = -30$
$b_0 = 0.01751$	0.25	YES $\|b_0\| < 0.25/2$	0.125	10	11	1	$B_0 = \text{ROUND}\ [b_0 \times 2^{10}] = 18$

The steps in computing the integer ROM coefficients for the serial method are as follows:

- *Step 1*: Set a temporary scale factor variable to SCALE = 1 and temporary bitwidth integer variable to $K = B - 1$. Apply the following quantization steps to the largest-magnitude original b_k floating-point coefficient (for example, b_4 in the upper left side of Figure 2).
- *Step 2*: If the b_k floating-point coefficient being quantized and all the remaining unquantized coefficients are less than the value SCALE/2, set SCALE = SCALE/2, set $K = K + 1$, and set the current coefficient's flag bit to Flag = 1. If the b_k floating-point coefficient being quantized or any of the remaining unquantized coefficients are equal to or greater than SCALE/2, variables SCALE and K remain unchanged, and set the current coefficient's flag bit to Flag = 0.
- *Step 3*: Multiply the b_k floating-point coefficient being quantized by 2^K and round the result to the nearest integer. That integer is our final value saved in ROM.
- *Step 4*: Repeat Steps 2 and 3 for all the remaining original unquantized b_k floating-point coefficients, in sequence from the remaining largest-magnitude to the remaining smallest-magnitude coefficient.

Table 1 illustrates the serial method quantization steps for the floating-point coefficients in Figure 2.

PARALLEL METHOD COEFFICIENT QUANTIZATION

In the parallel method, let's assume we want our ROM to store coefficients whose bitwidths are integer B. (For example, in the lower right side of Figure 5, $B = 8$.) Next, let's define an optimum magnitude range, R, as

$$0.5 \le R < 1. \qquad (1)$$

THE SO-CALLED SERIAL METHOD OF FILTERING IS COMPATIBLE WITH TRADITIONAL PROGRAMMABLE DSP CHIP AND FPGA PROCESSING.

The steps in computing the integer ROM coefficients for the parallel method are as follows:

- *Step 1*: Repeatedly multiply an original b_k floating-point coefficient (the upper left side of Figure 5) by two until the magnitude of the result resides in the optimum magnitude range R. Denote the number of necessary multiply-by-two operations as Q.
- *Step 2*: Multiply the original b_k floating-point coefficient by 2^{B+Q-1} (the minus one in the exponent accounts for the final coefficient's sign bit) and round the result to the nearest integer. That integer is our final value saved in ROM.
- *Step 3*: Repeat Steps 1 and 2 for all the remaining original b_k floating-point coefficients.

CONCLUSIONS

We introduced two novel methods for improving the precision of the fixed-point coefficients of FIR filters. Using the modified (compressed) coefficients, we achieved enhanced filter performance while maintaining phase linearity, without increasing the bitwidths of our filter multiplier or coefficients, nor the number of coefficients. The so-called serial method of filtering is compatible with traditional programmable DSP chip and FPGA processing, while the parallel method is most appropriate with an FPGA implementation.

ACKNOWLEDGMENTS

Many thanks to Rick Lyons for his careful analysis of these filtering methods and his patient assistance with the text of this article.

AUTHOR

Zhi Shen (zhi.m.shen@gmail.com) is pursuing the Ph.D degree with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China. As the project leader of the Digital TV Lab, he designed the multicarrier DVB modulator.

REFERENCES

[1] R. Lyons, *Understanding Digital Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2004, pp. 503–504.

[2] J. Proakis and D. Manolakis, *Digital Signal Processing—Principles, Algorithms, and Applications*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996, pp. 620–621.

[SP]

Al Wegener

Compression of Medical Sensor Data

Over the course of the last 20 years, compression has evolved from a somewhat esoteric domain of mathematics and computer science into a ubiquitous consumer electronics technology. Combined with steadily decreasing complementary metal–oxide–semiconductor (CMOS) silicon costs and increasing broadband Internet availability, compression algorithms for audio and video signals, such as Joint Photographer's Expert Group (JPEG), Moving Pictures Expert Group (MPEG), MPEG audio layer III (MP3), advanced audio coding (AAC), and H.264, have enabled the development of new consumer product categories such as digital cameras, digital audio players, digital versatile disc (DVD) players, high-definition television, and Internet videoconferencing. Without the 4x–32x decrease in effective bit rates that compression (and Moore's law) unlocked, these popular products would not be cost-effective and thus would not have reached consumer electronics penetration levels.

Given the prevalence and success of speech, audio, image, and video compression algorithms in consumer electronics, it is surprising that many high-speed digital signal processing (DSP) systems such as wireless infrastructure, radar processing, and medical imaging sensor subsystems have until recently not considered compression as an alternative for reducing data acquisition bandwidth and storage bottlenecks. In consumer electronics products, compression is typically used as a source coding solution to reduce the bit rate of a single media stream, before compressed audio and video packets enter the transport and storage infrastructure. In contrast, this article describes sensor compression for medical transducers that is an integral part of the high-speed DSP transport and storage infrastructure itself, and compressing and decompressing hundreds or thousands of sensor channels in real time. While we certainly acknowledge the benefits of compressing medical images that are output from image reconstruction, our goal is to describe the benefits of compressing medical sensor data before image reconstruction. When combined with low-cost, off-the-shelf computer and networking components, sensor compression reduces the costs of the medical imaging data acquisition, transport, and storage infrastructure. By reducing medical imaging equipment bill of material (BOM) costs, integrated medical sensor compression and decompression will ultimately reduce the cost of medical care.

SENSOR COMPRESSION, NOT IMAGE COMPRESSION

If sampled data compression has worked so well for speech, audio, image, and video, why haven't medical imaging engineers used compression before? In fact, medical image compression is seeing broad deployment in picture archiving and communication systems [1] that hospitals and doctors' offices use to manage the flood of medical images from all modalities. Compressing medical images after they are reconstructed has obvious benefits when the images must be transported to geographically distant doctors or when they are archived. Image compression algorithms such as JPEG2000 and JPEG-LS work well on the images that are output from image reconstruction. In contrast, this article describes the benefits of compression of sensor data that is the input to image reconstruction. Compression of medical sensor samples reduces medical imaging infrastructure transport and storage costs prior to image reconstruction.

There are several reasons why compression of medical sensor data prior to image reconstruction has not been deployed in medical imaging scanners. Medical sensor signals differ from audible and visual signals in three key aspects. First, medical imaging sensors generate hundreds or thousands of sensor streams, while audio and video signals are limited to at most six surround-sound channels or a few interleaved video streams. Because the typical number of sensor streams to be compressed is two or three orders of magnitude higher than the typical number of audio and video streams, the input–output (I/O) and processing architecture of medical sensor compression algorithms will differ substantially from the I/O and processing architecture of audio and video compressors. Second, the bandwidth and dynamic range requirements of medical sensors vary widely, from 10 ksamp/s at 20 b/sample for computed tomography (CT), to 100 Msamp/s at 16 b/sample for magnetic resonance imaging (MRI). In contrast, audio compression rates are determined by human hearing's unchanging 20 kHz bandwidth and 130 dB dynamic range, so audio decompression need not generate signals faster than 40 ksamp/s, or with more than 24 b/sample, or more than two channels (stereo). Third, the ultimate consumer of decompressed audio and video is a human with predictably limited hearing and vision, while the ultimate consumer

Digital Object Identifier 10.1109/MSP.2010.936781

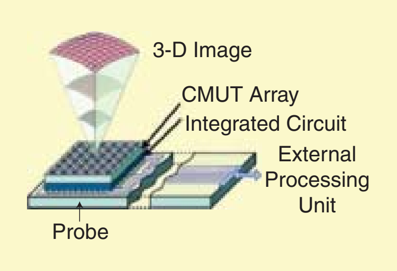
1053-5888/10/$26.00©2010IEEE

of decompressed medical sensor data is an ultrasound beamforming algorithm or a CT image reconstruction kernel. While the degradations introduced by audio and video compression algorithms are unnoticeable by humans, medical image reconstruction algorithms require the full bandwidth and dynamic range of the sensor data. Image reconstruction algorithms may generate unpredictable and unwanted artifacts if their input sensor data were subjected to compression techniques that were developed for human hearing or vision. To summarize, medical image compression algorithms support channel counts, sample rates, bit widths, and compression algorithms that differ fundamentally from those used by consumer audio and video systems.

MEDICAL SENSOR DATA RATES ARE RISING EXPONENTIALLY

All medical imaging modalities face the challenge of exponentially rising sensor data rates. As shown in Table 1, sensors in next-generation CT, ultrasound, MRI, and digital X-ray systems will generate from 4 Gb/s to 200 Gb/s of sampled data. Medical sensor data rates rise for three reasons. First, the number of sensor channels is increasing, allowing larger areas of anatomy to be scanned more quickly. Today's high-end CT scanners already contain over 300,000 X-ray scintillators, photodiodes, and analog-to-digital converters (ADCs), and that number typically doubles every three years. Second, sample rates per sensor channel are increasing. In medical ultrasound, yesterday's 3 MHz piezoelectric transducers are evolving to 18 MHz center frequencies, and the sample rates will rise proportionally [2]. Third, the dynamic range of medical imaging sensors is increasing. For instance, yesterday's 12-b ADCs for MRI are being replaced with 16-b ADCs because the wider sensor dynamic range improves MRI images. The combination of increasing channel counts, sample rates, and bit widths is the reason that medical imaging manufacturers are evaluating compression of medical sensor data to reduce bandwidth and storage costs.

ALL MEDICAL IMAGING MODALITIES FACE THE CHALLENGE OF EXPONENTIALLY RISING SENSOR DATA RATES.



[FIG1] Compression reduces 4-D ultrasound bandwidth requirements (from [9]). (Figure used with permission.)

Transporting and storing medical sensor samples can become expensive. As shown in Table 1, medical sensors generate Gb/s of data that must be transported, processed, stored, and displayed. Medical imaging scanners are often assembled from standardized, low-cost computer and networking components wherever possible because doing so lowers the cost of the scanners. Despite the low cost of these individual components, a next-generation CT scanner's sensor subsystem will generate 80 Gb/s of sensor data that must be delivered to the scanner's image reconstruction subsystem. CT scanners could use commercial gigabit Ethernet (GbE) links to transport the sensor data from gantry to image reconstruction, but this solution would require 80 or more GbE cables and 80 router ports on both ends of the link. Using 4:1 compression of the CT sensor data allows 20 GbE links to carry the same amount of data, reducing link costs by 75%. As illustrated in Figure 1, next-generation ultrasound machines will use two-dimensional (2-D) transducers that ultimately create a four-dimensional (4-D) image (a 3-D volume changing over time). To improve manufacturing yield, 2-D ultrasound transducers may use capacitive micro-machined ultrasound transducers (CMUTs) to replace piezoelectric crystals. Each of the 2,000 transducers in a 2-D probe must be sampled at 20+ Msamp/s with 10–12 b/sample, generating 200 Gb/s of sensor data that must be delivered to a beamformer and a scan converter (ultrasound's image reconstruction steps). Using 3:1 compression in the 2-D probe would reduce sensor and field-programmable gate array (FPGA) pin counts, power consumption, printed circuit board (PCB) complexity, and cabling costs by two-thirds. In summary, introducing compression into medical sensor transport and storage infrastructure reduces interconnect and storage costs by a factor of two to four, even when medical imaging scanners are already assembled using low-cost computer and networking components.

[TABLE 1] BANDWIDTH REQUIREMENTS FOR CT, ULTRASOUND, MRI, AND DIGITAL X-RAY SYSTEMS.

SENSOR SIGNAL TYPE	NEXT-GENERATION SENSOR BANDWIDTH	TYPICAL LOSSLESS* COMPRESSION RATIOS	FIXED-RATE* COMPRESSION RATIOS GENERATING CLINICALLY ACCEPTABLE IMAGES
COMPUTED TOMOGRAPHY	~80 Gb/s	2:1	≤5:1
ULTRASOUND (RF SIGNAL)	~200 Gb/s	1:85:1	≤3:1
ULTRASOUND (BEAMFORMED SIGNAL)	~20 Gb/s	2:1	≤4:1
MAGNETIC RESONANCE IMAGING	~5 Gb/s	4:1	≤6:1
DIGITAL X-RAY	~4 Gb/s	1:8:1	≤3:1

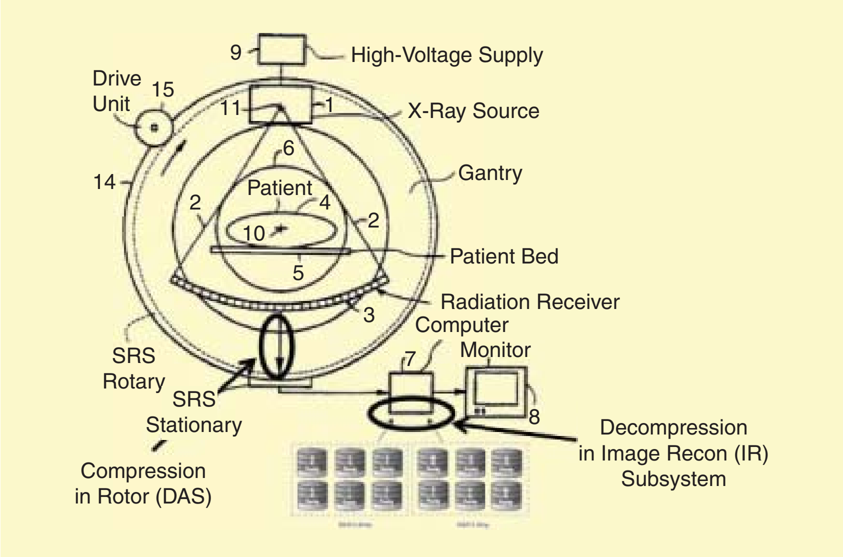
*The listed compression ratios were obtained using Samplify's Prism 3 compression on ultrasound, MR, and digital X-ray signals, and Prism CT compression for CT signals.

EXAMPLE: BENEFITS OF SENSOR COMPRESSION FOR CT

If medical imaging companies are to use compression, the benefits of doing so must be more cost-effective than just buying more bandwidth and storage. Let's consider a specific example: what would happen to the bill of materials (BOM) cost of a CT scanner if 3:1 compression could be achieved on CT sensor samples, instead of purchasing two to four times more CT bandwidth and storage? Figure 2 shows the block diagram of a CT scanner, which includes two primary functional components. The first component, a gantry, consists of a rotating platform that transmits rotating X-ray beams though a patient who lies on a table that moves thru the center of the gantry. The second component, a console, receives the digitized X-ray measurements and converts them into CT images using image reconstruction software. The gantry contains a data acquisition subsystem (DAS) with three types of elements that are replicated up to 300,000 times: scintillators that convert X-rays to light, photodiodes that convert light to current, and ADCs that convert current to digital samples. The gantry also contains a rotating, noncontacting electromechanical device called a slip ring, which transfers power from stator to rotor and sends digitized X-ray samples from the DAS on the rotor to the stator. The CT console contains a high-speed storage system [typically a redundant array of inexpensive disks, (RAID)], an image reconstruction compute fabric such as a multicore central processing unit (CPU), FPGA farm, or multiple graphics processing units (GPUs), and a display subsystem. CT sensor compression of 3:1 reduces slip ring costs by at least half. By reducing the slip ring bit rate by a factor of three, CT sensor compression also reduces console costs

- by decreasing the number of RAID arrays required to capture DAS samples in real time.
- by lowering the number of servers and RAID controllers associated with the RAID arrays.
- by reducing the physical size and power consumption of the console.

With a typical BOM cost of US$500,000 for a high-end CT scanner, sensor compression can reduce gantry costs by US$5,000 or more and console costs by US$25,000 or more. Even when accounting for compression's implementation costs, CT scanner BOM savings between 5–25% are the primary factor motivating CT scanner companies to evaluate sensor compression.



[FIG2] Compression reduces CT slip ring and storage array costs.

LOSSLESS COMPRESSION OF MEDICAL SENSOR DATA BECOMES ATTRACTIVE WHEN SENSOR SIGNAL COMPRESSION CAN HALVE THE SENSOR DATA RATE.

CHARACTERISTICS OF A REAL-TIME SENSOR COMPRESSION ALGORITHM

Given the wide variety of signals that medical imaging sensors generate, could one compression algorithm effectively compress all of them? Interestingly, medical sensor signals share three common characteristics: 1) they are usually slightly to moderately oversampled, 2) they have moderate to high peak-to-average ratios (PAR), and 3) they are sampled by imperfect ADCs. These characteristics are discussed in detail next.

The oversampling ratio is calculated by dividing the per-sensor sample rate by the sensor signal's effective bandwidth. The higher the oversampling ratio, the larger the sample-to-sample correlations and the smoother the signal waveforms will appear. Reducing sample-to-sample redundancy (a primary goal of compression) is relatively easy when medical sensor samples are moderately oversampled. Oversampling regularly occurs in DSP systems because sharp analog anti-alias filters prior to data conversion are more expensive than equivalent-rolloff filters in the digital domain.

PAR is defined as the maximum signal value divided by the mean signal value across an ensemble of signal samples (such as one patient scan), and expressed in decibels (dB). Because ultrasound signals are pulsed, we expect them to have a relatively high PAR of 8–10 dB. Ultrasound signals require their largest-magnitude samples for a very short time, typically at the start of a pulse. Because the amplitude of a received ultrasound signal quickly attenuates as it is reflected by various layers of blood, tissue, and bone, the mean ultrasound signal level is always 8–10 dB lower than its peak signal level. Surprisingly, CT and MRI sensor signals also exhibit PAR levels above 6 dB. Because PAR levels above 6 dB are common for many medical sensor signals, a compression algorithm employing a varying number of bits to encode signal amplitude provides a simple yet surprisingly effective compression technique.

Medical imaging equipment requires hundreds or thousands of ADCs to acquire sensor signals. ADCs and digital-to-analog converters (DACs) are characterized by two primary parameters: resolution (nominal number of bits per sample) and effective number of bits (ENOB). While resolution defines the numerical range of the ADC or DAC, ENOB is a measurement that defines the subset of the numerical range that contains useful information. An ADC with 14 b of resolution may only deliver 12.5 ENOB, where ENOB was measured using a full-scale sine wave. Because data converter ENOB values are always less than their resolution, some ADC and DAC samples are more useful ("effective") than others. An algorithm that preserves ENOB during compression will be useful for medical sensor signals.

Regarding the sometimes confusing term "real time," certainly a medical sensor compression algorithm must operate fast enough to compress all samples generated by all sensor ADCs. A second aspect of "real time" is related to the speed of image reconstruction. In CT and MRI, image reconstruction is usually a nonreal-time process. A CT scanner scans the patient in 30 s, but the CT images may only be available after ten minutes of image reconstruction. Under these circumstances, the compressed sensor data need only be decompressed as fast as samples are consumed by image reconstruction. A medical imaging scanner may thus have two "real-time" rates: a faster rate for sensor compression during data acquisition, and a slower rate for sensor decompression during image reconstruction.

Because the ADCs that digitize medical imaging sensor signals are usually attached to FPGAs, compression of sensor data will first be implemented in these existing FPGAs. However, because medical imaging often involves hundreds or thousands of sensor channels, compression's lowest cost and power consumption will only be realized when compression is integrated into sensor ADCs. By integrating compression into ADCs, the silicon area of compression is minimized, as are the pins, cables, and power used to transmit compressed packets from the ADC to subsequent signal processing. In both FPGA and ADC implementations of sensor compression, the compressor must be fast enough to compress all ADC sensor channels as fast as the ADC samples arrive.

Just as the implementation of medical sensor compression depends on the hardware available, implementation choices for decompression also depend on the available image reconstruction hardware. Certainly FPGAs have adequate resources for both real-time compression and decompression of sensor samples. Medical imaging modalities with real-time image reconstruction functions such as ultrasound beamforming will want to decompress the sensor data prior to, and possibly after, beamforming, using the beamformer FPGA. In medical imaging modalities where image reconstruction is slower than sensor acquisition, decompression is preferably done in the same processing hardware (GPU, CPU, or FPGA) as image reconstruction. Reference [3] describes a GPU implementation of the Prism CT decompression algorithm on an Nvidia GeForce GTX260, a 192-core GPU. GPUs and multicore CPUs are becoming the preferred medical image reconstruction platforms because of their scalability and algorithmic flexibility. For this reason, sensor decompression algorithms that are available on GPUs and CPUs, and that use a modest amount of MIPS when compared with image reconstruction MIPS, will be preferred.

HOW SHOULD THE MEDICAL IMAGING COMMUNITY JUDGE, AND CONSERVATIVELY SELECT, A LOSSY COMPRESSION RATIO FOR SENSOR SIGNALS?

LOSSLESS AND LOSSY COMPRESSION

Compression algorithms offer either lossless or lossy operation. Computer users regularly use lossless compression for e-mail attachments and file transfers. Lossless compression of medical images is also well known [4]. A drawback of lossless compression is its lower compression ratio when compared to lossy compression. Lossy compression is far more commonly used than lossless compression for audio and video. Lossy compression is sometimes called fixed-rate compression, because the guarantee of a fixed bit rate means that the decompressed samples usually differ from the original sensor signal samples. MP3 users are familiar with lossy compression's tradeoff: the higher the compression ratio, the larger the loss of audio fidelity. By selecting lossless or lossy compression for a given application, the compression user has made an implicit tradeoff between bit rate and quality. Lossless compression provides the best quality but achieves an unpredictable (and time-varying) amount of compression. Users cannot ask lossless file compression programs to achieve a specific compression ratio because the amount of compression achieved depends on the input and thus cannot be determined beforehand. Instead, lossless compression programs achieve as much compression as is consistent with the constraint that the decompressed data must be identical to the original data.

Lossless compression of medical sensor data becomes attractive when sensor signal compression can halve the sensor data rate. Halving the BOM cost is usually enough to justify a lossless sensor compression algorithm's various costs. Table 1 demonstrates that a compression algorithm called Prism achieves about 2:1 lossless compression on CT, MR, ultrasound, and digital X-ray sensor signals. In its lossy mode, Prism achieves between 3:1 and 6:1 compression on the same medical imaging sensor samples, where the resulting images have acceptable diagnostic quality as judged by radiologists, sonographers, and other medical imaging professionals (hereafter called image quality experts). Higher lossy compression ratios create greater BOM cost savings.

Patients and doctors may at first be reluctant, or at least somewhat nervous, about using medical imaging machines that compress sensor data using a lossy algorithm. MP3 listeners may not care if audio compression's distortions are occasionally unmasked, but it would be completely unacceptable if a radiologist misdiagnosed a patient because of lossy sensor data compression. Lossy compression of sensor signals must not change the diagnostic quality of reconstructed medical images from which image quality experts make clinical diagnoses. This tradeoff between bit rate and signal quality raises a key question: how should the medical imaging community judge, and conservatively select, a lossy compression ratio for sensor signals?

EFFECT OF LOSSY SENSOR COMPRESSION ON CLINICAL IMAGE QUALITY

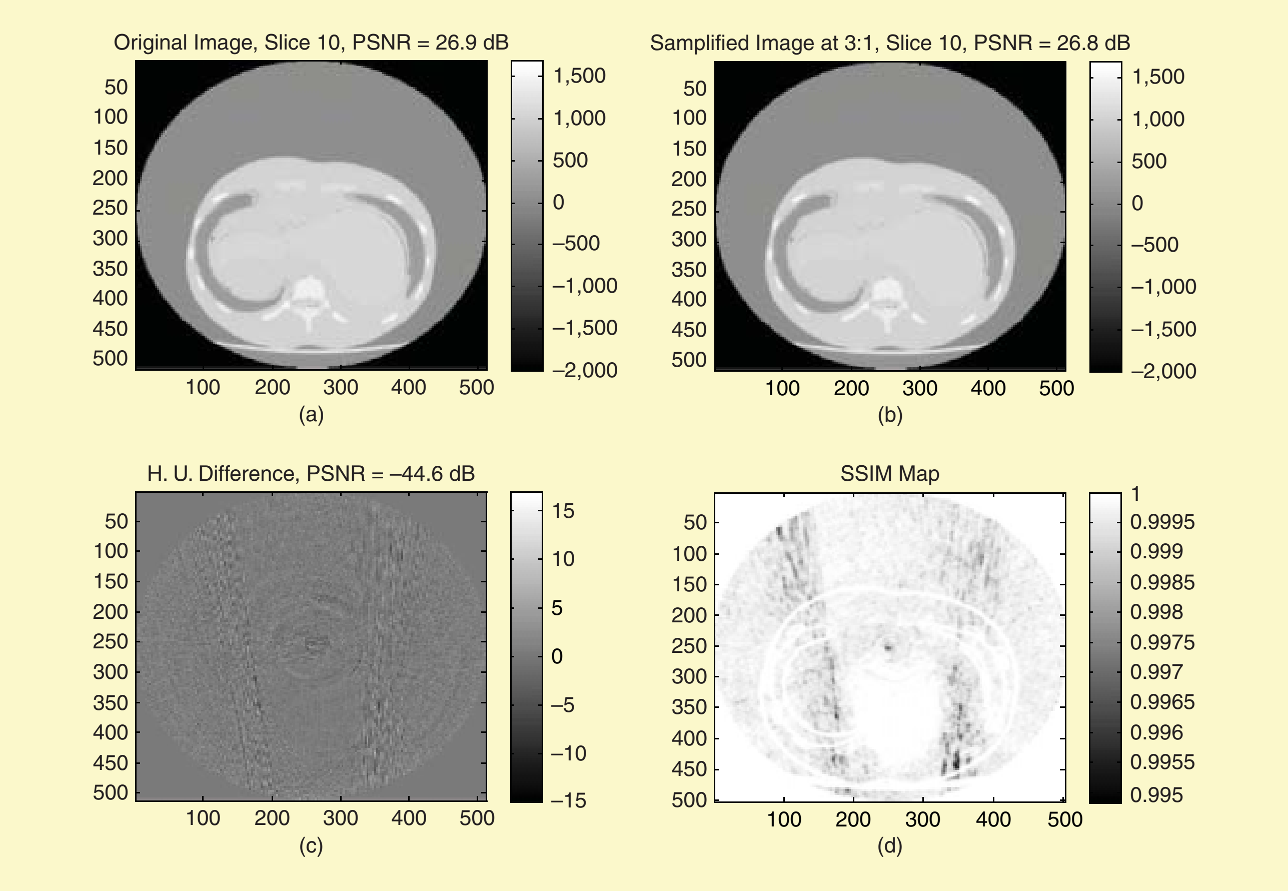
Appropriately designed medical image viewing tests reveal, with the overwhelming statistical strength that the medical imaging community requires, how much lossy compression can be applied to sensor samples before clinical image quality is reduced. Multiple image quality experts can be enrolled in receiver operating characteristic (ROC) tests that evaluate thousands of image pairs, where the image pairs include an original image (created from the original, noncompressed sensor signals) and an alternate image (created from compressed sensor signals at a variety of lossy compression ratios). A study in [5] describes a viewing test that used 1,890 CT image pairs from a commercial CT scanner, where alternate images were created from compressed CT sensor data at compression ratios up to 4:1. This study concluded that lossy compression of CT sensor data did not affect the clinical image quality on a variety of patient and phantom projection data sets, as judged both by a radiologist and several automated viewing metrics. Because such studies are new, we hope that our initial results motivate additional research into the effects of sensor compression on medical image quality. Since presenting the results of [5] the author has obtained additional sensor datasets (and the resulting images) that are ten times larger than those used in the original study. The involvement of additional radiologists would also be a welcome addition to this kind of research.

A medical sensor signal compression algorithm should provide a choice of compression ratios and then use the ratio that results in clinically acceptable images. Ideally the compression algorithm would also offer a lossless mode for those rare cases where unaltered, yet bit-reduced, sensor data is required. The chosen lossy compression ratio will vary, depending on modality (e.g., MR sensor signals compress more easily than ultrasound signals), anatomy (for example, head scans compress more easily than body scans), or function (for instance, real-time CT imaging for guided surgery at 30 frames/s acceptably has acceptably lower resolution than nonreal-time image reconstruction). Results of viewing studies that evaluate sensor compression's effects on image quality can be combined into a set of standardized scan protocols (parameter settings) that automatically select the appropriate compression settings for the variety of medical imaging tasks performed by today's scanners. Scan protocols ensure that patient images with acceptable clinical quality are delivered to doctors and radiologists for diagnosis, while patients, doctors, and hospitals benefit from less expensive medical imaging equipment that compression enables.

The image and video compression and machine vision communities have developed automated image quality assessment software that quantifies differences between image pairs. These automated tools are now being used to automatically assess medical image quality. While older metrics such as peak signal-to-noise ratio (PSNR) have been used for decades as a crude measure of digital image quality, image quality experts agree that PSNR is not well correlated with human perception of image quality. A relatively new image quality metric called structural similarity (SSIM) is both easy to calculate and well correlated with human image perception [6]. SSIM is applied to image pairs and measures their similarity on a scale from 0.0 (images are unrelated) to 1.0 (images are identical). In the previously described CT image quality study [5], SSIM was applied to the 1,890 CT image pairs whose alternate images were created using compressed CT sensor data. An SSIM threshold of 0.99 identified those images that were more likely to contain image artifacts that expert image viewers might notice. An SSIM threshold of 0.98 is often considered the level of visual indistinguishability. In Figure 3, (a) illustrates an original CT image, (b) shows an alternate image created from 3:1-compressed sensor data, (c) the pixel-by-pixel differences between the images, and (d) the SSIM map. Both the pixel-by-pixel differences and the SSIM map have been contrast-enhanced to the full grayscale range for easier perception of differences. An SSIM map sweeps a 10 × 10-pixel square across all possible 100-pixel regions of the original image and generates a local SSIM value for each region. The worst-case SSIM pixel of 0.995 in the SSIM map [Figure 3(d)] suggests that images in Figure 3(a) and (b) will look identical to expert viewers.

Automated image quality metrics such as SSIM can be used to identify medical images that might contain artifacts that image quality experts would notice. Images whose quality is above a predetermined SSIM threshold need not be examined by these image quality experts, because the image quality threshold was pre-calibrated using similar experts on many images. By using automated image quality metrics such as SSIM, image quality studies can include more images to increase confidence levels while keeping the study costs manageable.

SSIM is not the only automated metric for image quality. A metric called just-noticeable distortion (JND) that was originally developed to evaluate video compression standards is equally useful for medical imaging [7]. Because JND includes more perceptual variables, such as ambient lighting, display characteristics, and viewer distance from the display, JND's image quality estimates may ultimately be better-correlated than SSIM with human observer performance. While SSIM and JND provide comparable results [8], SSIM is freely available while JND software must be



[FIG3] **Comparing (a)–(b) two CT images (c) using pixel differences and (d) an SSIM map.**

licensed. SSIM is probably the preferred metric for automated image quality assessment, unless researchers already have access to the JND software.

FUTURE OF COMPRESSING MEDICAL SENSOR DATA

Compression for medical imaging sensor signals provides significant BOM cost savings and expands bandwidth-limited links. A compression algorithm that provides both user-selectable lossless and lossy compression modes, along with user-selectable compression ratios, can be fine-tuned to the desired rate-distortion tradeoff for medical imaging modalities that include CT, ultrasound, MRI, and digital X-ray. Image quality tests that utilize both automated image quality metrics and expert viewers will guarantee that lossy compression does not compromise clinical image quality. Compression can be enabled in FPGAs or directly in ADCs and DACs, while decompression is more often implemented in software on the same CPU or GPU platform that performs image reconstruction. Compression will become an integral part of next-generation medical imaging equipment while providing medical images with equivalent image quality to more expensive medical imaging systems that do not utilize compression. Just as compression has become an integral part of consumer electronics devices, compression of sensor samples is becoming an equally important function in medical imaging systems.

AUTHOR

Al Wegener (awegener@samplify.com) is the founder and CTO of Samplify Systems.

REFERENCES

[1] M. Zukoski, T. Boult, and T. Iyriboz, "A novel approach to medical image compression," *Int. J. Bioinform. Res. Applicat.*, vol. 2, no. 1, pp. 89–103, 2006.

[2] A. Evans, "Compression improves data bandwidth bottlenecks and costs in medical imaging systems," *Med. Design Technol.*, Aug. 2009 [Online]. Available: http://www.mdtmag.com/scripts/ShowPR~PUBCODE~046~ACCT~0007712~ISSUE~0908~RELTYPE~fe~PRODCODE~0860~PRODLETT~A.asp

[3] A. Wegener. (2009, Oct.). GPU-based decompression for medical imaging. Nvidia GPU Developer Summit [Online]. Available: http://www.nvidia.com/content/GTC/documents/1444_GTC09.pdf

[4] S. Miaou, F. Ke, and S. Chen, "A lossless compression method for medical image sequences using JPEG-LS and interframe coding," *IEEE Trans. Inform. Technol. Biomed.*, vol. 13, no. 5, pp. 818–821, Sept. 2009.

[5] A. Wegener, R. Senzig, N. Chandra, Y. Ling, and R. Herfkens, "Effects of fixed-rate CT projection data compression on perceived and measured CT image quality," in *Proc. SPIE Medical Imaging Conf.*, Feb. 17, 2010, vol. 7627.

[6] Z. Wang and A. Bovik, "Mean square error: love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[7] J. Johnston, E. Krupinski, H. Roehrig, J. Lubin, and J. Nafziger, "Human visual system modeling for selecting the optimal display for digital radiography," in *Computer-Assisted Radiology and Surgery Proc.*, 2004, vol. 1268, pp. 335–340.

[8] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[9] P. Khuri-Yakub, "3-D imaging using 2-D CMUT arrays with integrated electronics," Sep. 2003 [Online]. Available: http://www-kyg.stanford.edu/khuriyakub/opencms/en/research/ultrasonic/3D_Imaging/index.html

[SP]

Pew-Thian Yap, Guorong Wu,
and Dinggang Shen

Human Brain Connectomics: Networks, Techniques, and Applications

The human brain is organized into a collection of interacting networks with specialized functions to support various cognitive functions. The word "connectome" first burst on the scene with the work of Sporns et al. [1], who urged brain researchers to advance a comprehensive structural description of the elements and connections forming the human brain. An increasing body of evidence indicates that schizophrenia, multiple sclerosis, and autism exhibit abnormal brain connections. Changes in connectivity also appear to occur as a consequence of neuron degeneration, either from natural aging or diseases such as Alzheimer's disease. A connectome is hence fundamentally important for understanding brain growth, aging, and abnormality. At the micro level, the brain elements consist of single neurons, the amount of which often treads the realm of hundreds of billions, and possible connections between them numbering in the order of 10^{15}. At a more macro (and more manageable) level, the brain is parcellated into a number of regions, where each region accounts for the activity and coactivity of a population of neurons. The colossal task of constructing a connectome calls for powerful tools for handling the vast amount of information given by advanced imaging techniques. In this article, we provide an overview of the fundamental concepts involved, the necessary techniques, and applications to date.

NEUROIMAGING TECHNIQUES

In recent years, emerging magnetic resonance imaging (MRI) techniques with growing sophistication allow deeper insights beyond the brain's gross anatomy to probe functional connections. Functional MRI (fMRI), for example, capitalizes blood flow and oxygen consumption variations within the brain as markers for neuronal activity, and highlights brain circuits that are activated under different stimulated behaviors. Resting state fMRI (R-fMRI), detecting fluctuations in brain activity of a person at rest, can be employed to locate coordinated networks within the brain. High angular resolution diffusion imaging (HARDI) detects water diffusion along fibrous tissue and allows visualization of axonal bundles. The wealth of information provided by these imaging techniques furnishes new opportunities for in vivo investigation into brain circuitry.

THE BRAIN NETWORK

The N regions of a brain form the columns (targets) and rows (sources) of an $N \times N$ connection matrix C that may or may not be symmetric, depending on whether the connection directionality is important. The diagonal of the matrix is often zeroed since self-connectivity is not normally important in this context. The element c_{ij} of C represents connections between individual elements i and j. A confirmed absence of connection is denoted by a zero, while a confirmed presence of connection results in a one. A richer description of the connection is possible by adding physiological parameters, such as connection density, fiber length, and diffusion measurements, as additional layers of information. Combining these pieces of information then allows a structural description of both connection topology and biophysical properties. An illustration of the processes involved in constructing a brain network is given in Figure 1, with the details discussed in the upcoming sections.

BRAIN PARCELLATION

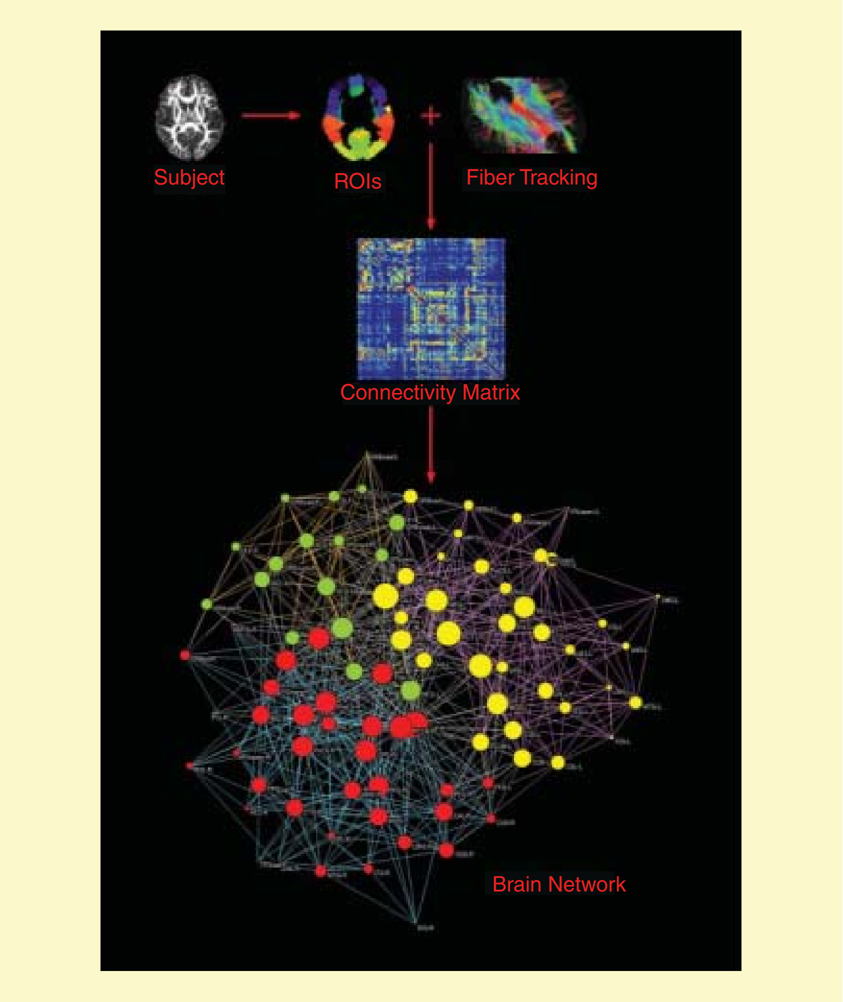
There are apparently a myriad of possible ways for parcellating a brain. There is, however, currently no single universally accepted parcellation scheme for human brain regions. Possibilities range from the commonly used modest 90 anatomically motivated parcellations given by the automatic anatomical labeling scheme [2], to the approximately 1,000 regions of interest defined at the white-gray matter interface used in Hagmann et al.'s work [3]. However, the question of whether structural parcellation of the brain results in functionally distinctive regions is largely unanswered. This is fundamentally important, since understanding brain function in relation to the structural substrate has been a major goal of neuroimaging. Perhaps the most promising approach is what termed as the connectivity-based parcellation, discussed by Behrens et al. in [4], where, based on the observation that functionally distinct gray matter regions manifest different patterns of remote connectivity, the gray matter is parcellated according to its connectional architecture, inferring boundaries between discrete functional regions.

REGISTRATION AND PARCELLATION PROPAGATION

Large-scale comparison of medical images for the purpose of studying brain connectivity cannot yet be performed without first removing confounding intra- or interindividual variations. Factors such as genetics, gender, pathologies, injury, and growth

Digital Object Identifier 10.1109/MSP.2010.936775

1053-5888/10/$26.00©2010IEEE



[FIG1] **Schematic illustration of the major processes involved in constructing a brain network using fiber tractography. A pair of regions are considered as connected if they are traversed by common fibers, and the numbers of connection fibers are recorded as elements in the connectivity matrix, which is then thresholded to retain only the significant connections. The nodes and intramodular connections are color coded for easier visualization of the communities detected via modularity maximization. The sizes of the vertices are weighted by the (logarithmically scaled) node betweenness.**

induce structural variations in the brain. Here, the importance of image preprocessing is therefore to align the population of images into a common space, matching spatially the structures in question. The increased specificity requirement in delineating connection abnormalities or growth related changes place increasing demands on registration algorithms. Thus, for the past two decades, we have seen a flourish of registration algorithms that cater for a wide range of imaging modalities. Upon establishing structural correspondence between a population of brain images and a brain atlas, parcellation information from the altas can be propagated to the individual images for consistent generation of connectivity matrices.

CONSTRUCTING THE CONNECTIVITY MATRIX

Different imaging modalities furnish complementary connectivity information. In what follows, we will discuss how connectivity is defined for some commonly used modalities. Constructing the connectivity matrix involves 1) gathering appropriate features from each region of interest and 2) establishing interregion correlation utilizing the gathered features. Details are as follows.

FEATURES

FUNCTIONAL CONNECTIVITY

FMRI measures the hemodynamic response related to neural activity in the brain and can be used to examine interregional correlation in neuronal variability. Regional functional connectivity is typically estimated using cross correlations, partial correlations, or mutual information of regional time series at one or several specific frequencies. The default mode network (DMN), for example, is characterized by coherent neuronal oscillations at a rate lower than 0.1 Hz.

STRUCTURAL CONNECTIVITY

The cerebral cortex is the outermost layer of neural tissue in the human cerebrum. It plays a key role in memory, attention, perception, thought, language, and consciousness. Connection networks can also be inferred from structural MRI data with brain regional connectivity estimated as correlations in cortical thickness [5] or volume [6]. After parcellating the brain into a number of regions, the mean cortical thickness or gray-matter volume are normally computed for the purpose of estimating the interregion connectivity.

WHITE-MATTER CONNECTIVITY

Diffusion weighted imaging (DWI) has gained considerable interest in the research community owing to its demonstrated capability of allowing in vivo probing of brain white-matter microstructures. In terms of characterizing crossing fibers, HARDI affords more information than the popular diffusion tensor imaging (DTI) and allows superior delineation of the angular microstructure of the brain white matter, making possible multiple-fiber modeling of each voxel for better characterization of brain connectivity. Fiber tractography allows the tracing of fiber bundles defined by the local maxima of the orientation distribution function of each voxel, and a pair of regions traversed by a significant amount of common fibers are considered as connected.

CONNECTIVITY MEASURES

Interregion dependence can be estimated with the help of correlation measures evaluated using fMRI time series, cortical thickness, or gray-matter volume. Pearson's correlation coefficient is commonly used for inferring connectivity from the measured feature values. It is defined as the ratio of the covariance of the two variables X and Y to the product of their standard deviations (σ_X, σ_Y)

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where μ_X and μ_Y are the means of X and Y, respectively. The computation of the partial correlation coefficient involves an additional step of regressing out the effect of a set of controlling variables, resulting in residuals from which the correlation coefficient can be computed. The controlling variables can include factors such as age, intracranial volume (ICV), or other sources of confounding covariances.

NETWORK ANALYSIS

We provide here formal definitions of some metrics commonly used in the analysis of brain networks. Representing a network as an unweighted graph G with N nodes, its metrics for global efficiency E_{glob} and local efficiency E_{loc} can be computed as

$$E_{\text{glob}} = \frac{1}{N}\sum_{i=1}^{N} E_{\text{glob}}(i),$$

$$E_{\text{glob}}(i) = \frac{1}{N-1}\sum_{\{j:j\neq i\in G\}} \frac{1}{l_{i,j}}$$

$$E_{\text{loc}} = \frac{1}{N}\sum_{i=1}^{N} E_{\text{loc}}(i),$$

$$E_{\text{loc}}(i) = \frac{1}{N_{G_i}(N_{G_i}-1)}\sum_{\{i',j':i'\neq j'\in G_i\}} \frac{1}{l_{i',j'}},$$

where $E_{\text{glob}}(i)$ and $E_{\text{loc}}(i)$ are nodal efficiency metrics, $l_{i,j}$ is the shortest path length between nodes i and j, G_i is a subgraph comprising nodes directly connected to node i, and N_{G_i} is the number of nodes of G_i. Specifically, E_{glob} measures the efficiency of parallel information transfer in the network, whereas E_{loc} measures the efficiency of local information transfer in the immediate neighborhood of each node.

A module of G is a subset of nodes that are more densely connected to each other in the same module than to nodes outsides the module. For a configuration of modular organization m with n_m modules, its modularity $Q(m)$ is defined as

$$Q(m) = \sum_{s=1}^{n_m}\left[\frac{h_s}{H} - \left(\frac{d_s}{2H}\right)^2\right],$$

where H is the total number of edges of G, h_s is the total number of edges in module s, and d_s is the sum of the degrees of the nodes in module s. The modularity of a graph is defined as the largest value of modularity measures associated with all possible configurations of modules, which can be found by optimization algorithms.

Betweenness measures the centrality of a node in a network, and, in some sense, indicates the influence of the node over the spread of information throughout the network. It is calculated as the fraction of shortest paths between node pairs that pass through the node of interest. The betweenness centrality of a node i, is defined as

$$B_c(i) = \sum_{j\neq k\neq i\in G} \frac{\sigma_{j,k}(i)}{\sigma_{j,k}},$$

where $\sigma_{j,k}$ is the number of shortest paths from node j to k, and $\sigma_{j,k}(i)$ is the number of shortest paths that traverse node i.

APPLICATIONS

Recent attempts of utilizing networks as a basis for understanding the brain at a "systems" level have brought new insights into the human brain, demonstrating the fact that a comprehensive description of the architecture of the anatomical connectivity patterns is fundamentally important in cognitive neuroscience and neuropsychology, as it reveals how functional brain states emerge from their underlying structural substrates and provides new mechanistic insights into the association of brain functional deficits with the underlying structural disruption.

SMALL WORLD NETWORKS

Recent research has reached a consensus that the brain manifests small-world topology, which implicates both global and local efficiencies at minimal wiring costs [7]. There are three classes of small-world networks: a) scale-free networks, characterized by a vertex connectivity distribution that decays as a power law; b) broad-scale networks, characterized by a connectivity distribution that has a power law regime followed by a sharp cutoff; and c) single-scale networks, characterized by a connectivity distribution with a fast (Gaussian or exponential) decaying tail. Each network has different degree of resilience to targeted attacks. Studies have also indicated that various neurological diseases, such as Alzheimer's disease [8], schizophrenia [9] and multiple sclerosis [5], cause disruption in the small-worldness nature of the networks. In [8], for instance, the clustering coefficients for the left and right hippocampus were found to be significantly reduced in a Alzheimer's disease group compared to a control group.

CLASSIFICATION AND IDENTIFYING POPULATION DIFFERENCES

Research has moved on to utilize whole-brain connectivity information as the basis for classification and locating population regional differences. In Robinson et al.'s work [10], for example, pattern features were extracted from the connectivity matrices of two age groups (20–30 and 60–90 years) using principal component analysis (PCA) and linear discriminant analysis (LDA). Employing these features for classifying subjects from these two age groups, a K-fold cross validation indicates that a mean accuracy of 87% can be achieved, indicating significant connection changes with aging. In the same framework, they have further identified the key differences between these two age groups.

CONCLUSION

A description of human brain connectome is important for the understanding of brain neurological function, development, and disease mechanism. Effort in this direction can be conducive to diagnosis and the identification of possible biomarkers of neuropsychiatric disorders. This article introduces the fundamental concepts involved in constructing a human brain connectome, commonly used techniques, and some applications to date. The construction of brain connectome will provide a new and exciting venue for the application of signal processing techniques in medical imaging.

ACKNOWLEDGMENT

This work was supported in part by NIH grants EB006733, EB008374, EB009634, and MH088520.

AUTHORS

Pew-Thian Yap (ptyap@med.unc.edu) is with the University of North Carolina at Chapel Hill as a postdoctoral research associate.

Guorong Wu (grwu@med.unc.edu) is with the University of North Carolina at Chapel Hill as a postdoctoral research associate.

Dinggang Shen (dgshen@med.unc.edu) is with the University of North Carolina at Chapel Hill, where he is an associate professor of radiology, BRIC, computer science, and biomedical engineering.

REFERENCES

[1] O. Sporns, G. Tononi, and R. Kotter, "The human connectome: A structural description of the human brains," *PLoS Comput. Biol.*, vol. 1, no. 4, pp. 245–251, 2005.

[2] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[3] P. Hagmann, M. Kurant, X. Gigandet, P. Thiran, V. J. Wedeen, R. Meuli, and J.-P. Thiran, "Mapping human whole-brain structural networks with diffusion MRI," PLoS One, vol. 2, no. 7, pp. 1–9 (e597), 2007.

[4] T. E. J. Behrens and H. Johansen-Berg, "Relating connectional architecture to grey matter function using diffusion imaging," *Philos. Trans. R. Soc. B*, vol. 360, no. 1457, pp. 903–911, 2005.

[5] Y. He, A. Dagher, Z. Chen, A. Charil, A. Zijdenbos, K. Worsley, and A. Evans, "Impaired small-world efficiency in structural cortical networks in multiple sclerosis associated with white matter lesion load," *Brain*, vol. 132, pt. 12, pp. 3366–3379, 2009.

[6] D. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, K. Q. Weinberger, and A. Meyer-Lindenberg, "Hierarchical organization of human cortical networks in health and schizophrenia," *J. Neurosci.*, vol. 28, no. 37, pp. 9239–9248, 2008.

[7] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[8] K. Supekar, V. Menon, D. Rubin, M. Musen, and G. Michael, "Network analysis of intrinsic functional brain connectivity in Alzheimer's disease," *PLOS Comput. Biol.*, vol. 4, no. 6, pp. 1–11 (e1000100), 2008.

[9] Y. Liu, M. Liang, Y. Zhou, Y. He, Y. Hao, M. Song, C. Yu, H. Liu, Z. Liu, and T. Jiang, "Disrupted small-world networks in schizophrenia," *Brain*, vol. 131, no. 4, pp. 945–961, 2008.

[10] E. C. Robinson, A. Hammers, A. Ericsson, A. D. Edwards, and D. Rueckert, "Identifying population differences in whole-brain structural networks: A machine learning approach," *Neuroimage*, vol. 50, no. 3, pp. 910–919, 2010.

continued from page 10

[TABLE 1] WORLDWIDE MEDICAL IMAGING SEMICONDUCTOR REVENUE FORECAST BY PRODUCT.

$M	2006	2007	2008	2009	2010	2011	2012	2013	2014	09–14 CAGR%
DSP	28.8	30.9	31.4	25.8	33.5	39.2	41.2	51	60.6	18.6%

Source: Databeans Estimates

The leading players in the medical ultrasound market, according to Global Industry Analysts, are Aloka Company, B-K Medical, Esaote SPA, GE Healthcare, Hitachi Medical Systems America, Medison Co. Ltd., Philips Healthcare, Siemens Healthcare, SonoSite, TomTec Imaging Systems GmbH, and Toshiba Medical Systems. Philips, Siemens, GE, and Toshiba reportedly account for about 80% of the global market.

The MRI market is projected to reach US$5.5 billion in 2010, driven by the introduction of high-field systems and new techniques such as functional neuro imaging, magnetic resonance angiography, noninvasive colonoscopy, and breast MR.

The key selling point in MRI device selection seems to be its high image quality and cost effectiveness. GE Healthcare, Siemens Medical Solutions, and Philips Medical Systems dominate the global MRI equipment market, according to Global Industry Analysts, while other prominent players include Esaote, Hitachi, Toshiba Medical Systems, Fonar Corp., IMRIS, and Medtronic.

Frost & Sullivan, another research organization that studies the medical imaging market, recently published a report suggesting there is a flurry of research and development (R&D) activity in medical imaging in Europe, particularly for cardiology applications. F&S anticipates a significant market opportunity in the echocardiography segment for manufacturers that can offer portable, PC-based ultrasound systems to private practitioners.

BIG DSP REQUIREMENT

Where does digital signal processing fit into the medical systems market?

Databeans is projecting that revenue from DSPs sold into worldwide medical imaging applications will nearly double from US$31.4 million in 2008 to US$60.6 million in 2014 (see Table 1) as the market for these systems grows and the technology advances on several fronts.

One technical advancement has been the migration of X-rays from film to digital files. DSP is helping convert X-ray signals to digital images at the point of acquisition, with no tradeoffs in image clarity. As TI notes in a report on the future of medical imaging, the ability to render digital images in real-time has led to the use of digital X-ray machines in surgical procedures, enabling doctors to view a precise image during surgery.

MRI is also improving with higher quality images in a fraction of the time required just a few years ago. Also, diffusion MRIs allow researchers to create brain maps to study the relationships between disparate brand regions via tractography. Functional MRIs can now rapidly scan the brain to measure signal changes caused by changing neural activity. DSPs are also playing a key role in telemedicine, particularly in videoconferencing and telepresence systems to support a variety of codecs.

The use of DSP is "a common theme that flows through all of these examples," according to the TI report. More importantly, the technology is having a major impact on healthcare worldwide.

Robert C. Streijl, Stefan Winkler,
and David S. Hands

Perceptual Quality Measurement—Towards a More Efficient Process for Validating Objective Models

The Quality of Service Metrics (QoSM) Committee of the Alliance for Telecommunication Industry Standards (ATIS) Internet Protocol Television (IPTV) Interoperability Forum (IIF) is tasked with defining how objective quality metrics can provide meaningful IPTV performance measures. This group has reviewed current objective quality models as well as the processes by which such models are validated. This article describes current practices in validating objective quality models and presents a new, streamlined process that can be implemented to achieve more efficient and effective model validation. Of main interest for IPTV are models for predicting video and audiovisual quality; however, the process also applies to the validation of perceptual quality models (PQMs) for other modalities. The proposed process offers vendors a fast route to validating objective PQMs while providing industry with the assurance of independent, unbiased model evaluation.

BACKGROUND

Service providers are rolling out IPTV services to slow down erosion of revenues from circuit-switched voice services and to keep up with the competition to deliver multiplay service offerings. To support IPTV operations, the need for service performance measurements that can provide insights into the customer's perception of the quality of IPTV content is apparent. Vendors, standards groups, and researchers are actively investigating meaningful algorithms and tools for conducting these measurements.

Subjective quality tests are widely used and support the development and testing of objective perceptual quality models (or objective models) that predict customer perception as a benchmark. However, subjective quality tests are not a practical solution for in-service performance monitoring. The purpose of objective models is to replace subjective tests by estimating the perceptual quality of voice, audio, video, and multimedia. Objective models can use different techniques to predict subjective quality. These techniques include full-, reduced- and no-reference methods and may utilize pixel-domain, bit stream, packet data, or some combination of these information sources to extract parameter values that then are used to predict quality [1]–[3]. The industry has an increased need for objective models as competition increases, and as quality becomes both a critical part of the value chain [e.g., high-definition TV (HDTV)] and a potential market differentiator between service providers.

THE MAIN PREMISE IS THAT AN OBJECTIVE MODEL ALGORITHM DOES NOT NEED TO BE STANDARDIZED IN ITSELF, AS ITS PRIMARY REQUIREMENT IS MEASURING QUALITY WITH A CERTAIN LEVEL OF ACCURACY.

The creation of objective models to compute an estimated customer opinion score is a complex process. Fundamental to the success of objective models is how accurately they can predict subjective quality ratings. A set of statistical methods has been defined to determine the accuracy of objective models [9].

The accuracy of objective PQMs is currently validated through various routes, including self-validation, contracted external validation, and independent validation (e.g., by the Video Quality Experts Group (VQEG) [13]). Clearly, the industry will find great value in model accuracy data that is obtained through independent validation, as well as data that is based on appropriate subjective testing methods and model performance metrics.

This column considers the limitations of current validation procedures, such as those practiced by VQEG, ITU-T Study Group 9 (SG9), and ITU-T Study Group 12 (SG12), presents work in progress within relevant standards groups (in particular the ATIS IIF QoSM Committee) to address these problems, and outlines a proposal for providing more effective and efficient model validation.

CURRENT VALIDATION PROCESSES: VQEG AND ITU

VQEG [13], [14] has been central to coordinating efforts to perform independent validation of objective perceptual quality models in a competition-style process. VQEG has completed several phases of testing to date, and the model performance data obtained from these tests has been used by the ITU to produce international standards [7], [8], [10]–[12]. The VQEG process is based on voluntary contributions from government organizations, research centers, universities, and industry. For agreed projects, VQEG prepares a test plan, in collaboration with those who participate, that defines the scope of testing, the types of objective models that may be submitted, subjective test methods and test laboratories that

Digital Object Identifier 10.1109/MSP.2010.936776

1053-5888/10/$26.00©2010IEEE

may perform subjective tests, model evaluation criteria, and so on.

The current VQEG process has the advantage of bringing together the premier experts in objective and subjective assessment to perform independent validation of objective models. Unfortunately, the relatively slow progress of VQEG projects means that the validation of models does not keep pace with industry requirements, and standardized models become outdated. The test plans often take several years to define, and once they are agreed upon, the test phase itself (including the accumulation of suitable test content, preparation of test sequences, and completion of subjective tests) is a lengthy process. After project completion, the best performing models may be standardized, and VQEG then moves on to the next project. The approach adopted by VQEG has the consequence that once a particular form of objective model has been validated and subsequently standardized, it may take many years before the group is able to perform a second validation test for that form of model. In fact, to date VQEG has not run a second validation round for any form of model. For example, full-reference TV (FR-TV) models were validated by VQEG in 2003 and standardized by the ITU in 2004. These models remain the standard so far as no further FR-TV validation tests have been performed, yet superior models may well have been developed in the meantime. The FR-TV test in particular did not include H.264 compression artifacts or IP loss impairments; consequently, the current standardized models have not been tested for the conditions that are present in most of today's IPTV systems.

Until recently, the VQEG process used Independent Test Laboratories (ITL) to perform subjective testing and model validation. More recently, VQEG has allowed model developers to act as test laboratories. This has led to a move away from cleanly separating the model development from the model validation. VQEG has begun working on an alternative process to validating "competing" models, having initiated a Joint Effort Group (JEG) that will test dedicated model components with the goal of building a model that combines the best performing modules from different organizations. Similarly, ITU-T SG12 has started a series of collaborative projects directed towards producing "best-of-breed" objective models. The approach taken by SG12 is to develop alternative objective models collaboratively that are then validated by the group. It should be noted that in the SG12 projects, many organizations that contribute objective models also perform the subjective tests used to validate the models and/or model components.

[VQEG HAS BEEN CENTRAL TO THE INDEPENDENT VALIDATION OF PQMs. UNFORTUNATELY, ITS RELATIVELY SLOW PROGRESS DOES NOT KEEP PACE WITH INDUSTRY REQUIREMENTS, AND STANDARDIZED MODELS BECOME OUTDATED.]

Reviewing the approaches of VQEG and SG12, several limitations in the current validation processes can be identified:

- Validating PQM models requires the acquisition of suitable multimedia content. Once this test material has been made available to model developers, it cannot be reused in future validation tests, requiring the selection and preparation of new content for subsequent tests.
- The current approaches (competition, collaboration, etc.) have strict cutoff dates for model submission, because all models are evaluated in the same exercise.
- Model developers are sometimes involved in the preparation of processed video sequences or in conducting subjective experiments due to ITL budget and time constraints, which is not ideal for an independent evaluation.
- At this time, the entire process for validating PQMs is very lengthy and can take several years, because a new test plan is written and a new test library is created for every round of testing.
- Once a standard has been defined and approved, it is very difficult to change, which means that standardized models can quickly become outdated, and there is no process for the models or the standards to be updated in a prompt fashion.

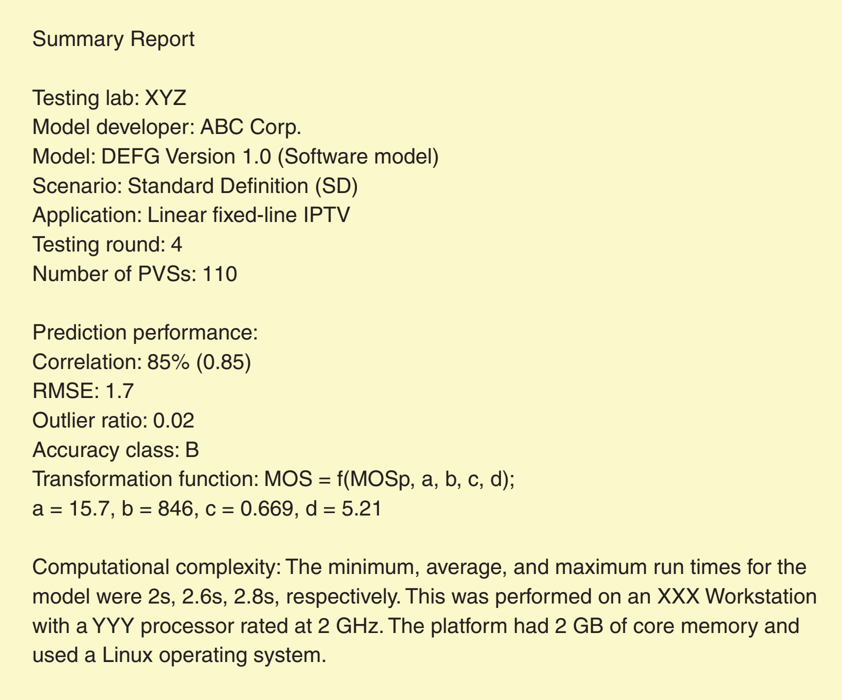
A NEW VALIDATION PROCESS

The ATIS IIF QoSM Committee has been working on a series of documents that form the backbone to validating objective models.

A general test plan for performing validation tests [4] was standardized to encourage industry developments where multiple organizations could develop PQMs all using the same basic test plan. With such a test plan in place, additional specialized documents, specific for each type of model, would then need to be developed that go into more detail for particular types of PQMs and applications. It is recommended that for each type of model, a single test plan is produced so that multiple organizations that want to test such a model all use the same procedures.

A technical report proposing a new process for validating objective models [5] has recently been completed. To date, standards groups combine the test process and test plan activities with the eventual goal of a standardized PQM solution. ATIS IIF separates these two processes. This column describes the concepts specified in the ATIS technical report.

Completing the series, a third document is planned that specifies the various types of perceptual quality measurements for use in IPTV environments [6]. The purpose of that document is to recommend a variety of IPTV quality of experience (QoE) measurements that predict customer experience, to describe the various types of measurements (e.g., parametric and bitstream approaches), their inputs and outputs, and also the points in an IPTV system where such measurements could be most useful.



Summary Report

Testing lab: XYZ
Model developer: ABC Corp.
Model: DEFG Version 1.0 (Software model)
Scenario: Standard Definition (SD)
Application: Linear fixed-line IPTV
Testing round: 4
Number of PVSs: 110

Prediction performance:
Correlation: 85% (0.85)
RMSE: 1.7
Outlier ratio: 0.02
Accuracy class: B
Transformation function: MOS = f(MOSp, a, b, c, d);
a = 15.7, b = 846, c = 0.669, d = 5.21

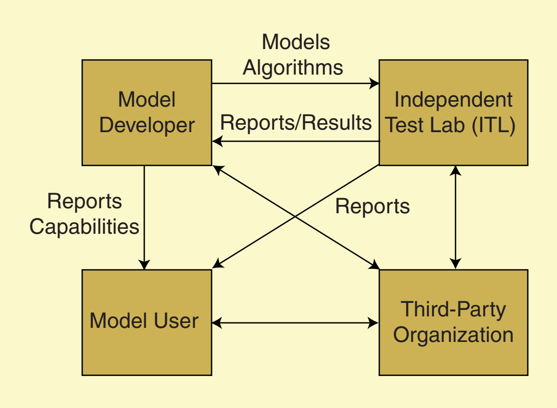
Computational complexity: The minimum, average, and maximum run times for the model were 2s, 2.6s, 2.8s, respectively. This was performed on an XXX Workstation with a YYY processor rated at 2 GHz. The platform had 2 GB of core memory and used a Linux operating system.

[FIG1] **Example summary report [5].**

The main premise is that an objective model does not need to be standardized in itself, as its primary requirement is measuring quality with a certain level of accuracy. Also, one could specify various types of perceptual quality models by e.g., their type, expected behavior, inputs, and outputs, thus allowing a black-box approach where the internal details of algorithms do not need to be revealed. Instead, with the test process, test plan, and specification of various types of perceptual quality measurement standardized, a repeatable process for model validation and comparison is created.

Given this, and considering the strengths and weaknesses of VQEG and SG12 approaches to model validation, the ATIS IIF QoSM Committee has produced an alternative process that is similar to the current processes in several ways but is believed to strengthen their weaker aspects. This process has the following unique characteristics:

- An independent validation process using a secret content library of video sequences annotated with subjective ratings, allowing content to be used more than once. The library is prepared and maintained by the ITLs; model developers only have access to the information that is publicly available to everybody and do not become involved in video creation or subjective testing in various ways. Because the library is designed to be reusable, it can be bigger and more varied than for a single test.
- On-demand algorithm validation that allows model developers to have a model evaluated at any time, e.g., at the request of a customer, or when a new model version is released.



[FIG2] **Participants of the test process [5].**

- Quick turn-around times for model validation rather than multiyear testing events. This is possible because the test procedures and annotated content libraries are prepared in advance, and checking model performance is a simple matter of running a model on the video sequences in the library and compiling results, something that can be done within a few weeks.
- Spur ongoing development and rapid improvement of models, thus increasing model quality and accelerating availability of the best models for model users.
- Clear, well-defined reporting templates, which are designed to provide an overview of the performance of a given model, as well as to facilitate easy comparison of multiple models. Model reports can be requested by model users from model developers or the ITLs. An example report is shown in Figure 1.
- Supporting these process improvements, a validation process is required that consists of clearly defined entities and entity roles, focused on single algorithm submission rather than processes based on competition or collaboration specifically. Collaboratively created models would be validated in the same way as a single algorithm.
- Only the model performance with respect to a standard test plan and library are published. There is no need for algorithm standardization as such. Model developers can keep the details of their algorithms secret, if they so choose, and license their models on their own terms. For example, a model can be developed for a single customer, who can still benefit from independent evaluation.

Other aspects may be quite similar to the current processes. It is envisioned that the process is open and could be as "democratic" in nature as the current processes. To initiate the process, ITLs, model developers, model users, and standards bodies should work together to define the scope and categories under consideration for model validation.

The validation process is composed of four building blocks (see Figure 2). The blocks represent the different parties needed to provide a rigorous and systematic approach to independently validating objective models.

Fundamental to the process is the existence of an ITL. The ITL may comprise one or more test laboratories and cannot develop objective models. The ITL's operations would be coordinated by a third-party organization that would be the overall sponsor of process activities as well as the business aspects of the process (e.g., relation with content providers, facilitate democratic participation of all parties in this process, fee schedules, and media communications). This third-party organization could be a nonstandards (e.g., commercial) entity or an international standards body such as ATIS or ITU.

The ITL would produce an extensive library of test sequences that are annotated with subjective quality ratings. The library of test sequences needs to be sufficiently large and representative of different video oriented services (e.g., HDTV, mobile) for it to be a good test of model performance. Furthermore, the library of test sequences must be secret. By possessing a large, secret library of test sequences, the ITL is able to reuse test materials for validating models. The ITL is expected to maintain and extend the library of test sequences over time, increasing existing data sets and creating new libraries to accommodate technology developments (e.g., new codecs). The library of test content should be representative of different content genres and should be designed with possible different model categories in mind (e.g., linear broadband TV versus wireless TV). A publicly available document providing a written description of the test content will be produced by the ITL. This written record of test content should provide a description of the video and, where appropriate, audio component of each test sequence.

Once the sequence library has been prepared, the ITL conducts subjective tests on the sequences in the library for annotation with mean opinion scores (MOS). Subjective scores will be obtained in line with the appropriate standardized subjective test procedures. The MOS annotations need to be maintained and extended along with the sequence library.

Once the annotated sequence library is in place, model developers can submit their models to the ITL for validation. The ITL will perform the validation tests by running the model against a large set of secret sequences that meet the defined scope of the tests.

Once completed, the ITL prepares a report that details the scope of the validation test and the performance of the model. The report is sent to the model developer, who can then decide whether or not to publicly release the performance data. The summary report using a well-defined template will allow model users to compare results from different models and choose the one best suited to their needs. The summary report (see Figure 1) includes reference to the test plan, category/service scenario/application tested, sequence library, and the number of sequences used in the validation test. It also specifies the prediction performance of the model for the set of PVSs in terms of evaluation criteria, such as correlation coefficients, prediction error, or outliers. Finally, the report includes some indications of model complexity and runtime.

THERE DOES NOT NEED TO BE A STANDARDIZATION COMPONENT FOR OBJECTIVE MODELS AS LONG AS THERE IS A RELIABLE INDEPENDENT VALIDATION PROCESS.

To compare PQMs and PQM results from different model developers, especially as multiple different solutions could be used in an operational environment, there is a need to translate (or cross calibrate) the output of one model with that of another. Cross calibration is a transformation of model outputs to a common scale through the annotated PVS database, typically using a linear or nonlinear fitting function that maps the MOS model outputs to the subjective MOS [15]. Computing this fitting function for a model is part of the validation and will be done by the ITL; the function and its coefficients will also be given in the summary report [5].

CONCLUSIONS

We described the shortcomings of current standards-based test processes for evaluating the accuracy of objective models. Based on the work of the ATIS IIF QoSM Committee, we introduced an improved process that mitigates the weaker points of the current processes. We also indicated that there does not need to be a standardization component for objective models as long as there is a reliable independent validation process.

The next step is to actually put this process in place. Practical and commercial questions need to be addressed, for example:

- Who are the ITLs?
- Who is the third-party organization?
- What is the fee structure for model validation?
- What is the role of VQEG, ITU, and ATIS in this process, if any?

This is part of an ongoing discussion among various standards groups, including ATIS IIF, VQEG, ITU-T SG9 and SG12.

RESOURCES

ATIS RESOURCES

The ATIS Web site (www.atis.org) provides details of ATIS standards and technical reports. Contributions to the ATIS IIF QoSM Committee are available from www.atis.org/IIF/.

VQEG RESOURCES

The VQEG Web site (www.vqeg.org) provides information on its past and present test projects. The test plans and test reports for each project are available for download. Communications between

standards in a NUTSHELL continued

ATIS IIF QoSM and VQEG can be found under "Meeting Files" for the various VQEG meetings.

ITU RESOURCES

The ITU Web site (www.itu.int) has links to all ITU-T and ITU-R publications. ITU members can access working documents including the test plans for validating parametric models currently under investigation by Study Group 12.

AUTHORS

Robert C. Streijl (robert.streijl@att.com) is a principal member of technical staff in AT&T's architecture and planning organization. He is the cochair of the ATIS IIF QoS Metrics Committee.

Stefan Winkler (swinkler@cheetahtech.com) is chief scientist at Cheetah Technologies. He is an active contributor to VQEG and ATIS IIF and cochair of the QoE Metrics Activity Group of the Video Services Forum.

David S. Hands (david.2.hands@bt.com) is a research group leader with BT Innovate & Design. He is an active member of ATIS IIF QoSM, ITU-T SG9, and VQEG standards groups.

REFERENCES

[1] D. Hands. (2007, Mar. 9–10). Video quality measurement: Past, present and future. *Proc. IMQA 2007*, Chiba Univ., Chiba, Japan [Online]. Available: http://www.mi.tj.chiba-u.jp/IMQA2007/

[2] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sept. 2008.

[3] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process. Image Commun.* (Special Issue on Image and Video Quality Assessment), to be published.

[4] ATIS, "Test plan for evaluation of quality models for IPTV services," ATIS-0800025, Oct. 27, 2009.

[5] ATIS, "Validation process for IPTV perceptual quality measurements," ATIS-0800035, Tech. Rep., Dec. 28, 2009.

[6] *QoE Measurement Recommendations and Framework*, ATIS-0800031, work in progress.

[7] *Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference*, ITU-R Recommendation BT.1683, June 2004.

[8] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, ITU-T Recommendation J.144, Mar. 2004.

[9] *Method for Specifying Accuracy and Cross-Calibration of Video Quality Metrics (VQM)*, ITU-T Recommendation J.149, Mar. 2004.

[10] *Perceptual Visual Quality Measurement Techniques for Multimedia Services Over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, ITU-T Recommendation J.246, Aug. 2008.

[11] *Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference*, ITU-T Recommendation J.247, Aug. 2008.

[12] *Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Reduced Reference*, ITU-T Recommendation J.249, Jan. 2010.

[13] Video Quality Experts Group (VQEG) official Web site [Online]. Available: http://www.vqeg.org/

[14] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," *IEEE Signal Processing Mag.*, vol. 26, no. 3, pp. 96–101, May 2009.

[15] ATIS, "Methodological framework for specifying accuracy and crosscalibration of video quality metrics," ATIS Tech. Rep. T1.TR.72-2001, Oct. 2001.

from the GUEST EDITORS

continued from page 12

The fourth article in this issue, by Pham et al., describes how digital topology is used to compute mathematical representations of the brain's complex and varied structures. Such methods are central to mapping the brain and can help to model global connectivity.

The most sophisticated of today's medical imaging techniques are based on tomographic reconstruction, a general approach in which images of the body's interior are computed from numerous images acquired from outside the body. Tomographic reconstruction is an inverse problem, in which the goal is to invert a sometimes complicated system describing the physical process of data acquisition. Some of the basic concepts of tomography date back to 1917, when Johann Radon described a formalism now known as the Radon transform. Yet, in spite of decades-long interest in the problem of reconstructing medical images, the past few years have seen an explosion of new discoveries about the nature of this inverse problem and its solution.

The fifth article in this issue, by Clackdoyle and Defrise, discusses dramatic recent developments in the solution of the tomographic image reconstruction problems, overturning long-held notions about fundamental issues in this problem domain. In particular, the article reviews advances with respect to reconstruction from incomplete data, and the so-called "interior problem."

Next, Fessler describes so-called model-based approaches to reconstruction in magnetic resonance imaging (MRI), an alternative to classical approaches based on direct Fourier inversion. These approaches recognize the complex nature of real-life MRI data, which include, for example, non-Fourier physical effects and nonlinear magnetic fields. In addition, these approaches can accommodate deliberate undersampling schemes adopted to permit fast scanning; thus, this work relates also to the field of compressive sensing, which was the subject of a prior issue of *IEEE Signal Processing Magazine*.

Finally, this issue concludes with a article by Ying and Liang, which discusses parallel MRI, an approach in which a phased array of coils is used to perform MRI more rapidly than traditional methods. Parallel MRI is a cutting-edge technology in medical imaging in which signal processing plays a central role. This article focuses on the signal processing issues of multichannel sampling and filter-bank theory.

A WORD OF THANKS

We were pleased and overwhelmed by the large number of outstanding articles submitted for consideration in this issue, and we regret that we could not accommodate all of them. We are very grateful to the authors and reviewers for their exceptional efforts and thoughtful contributions. We also thank Area Editor Dan Schonfeld for his valuable support and assistance throughout the process.

[best of THE WEB]

Alessandro Vinciarelli
and Maja Pantic

Techware: www.sspnet.eu: A Web Portal for Social Signal Processing

Please send suggestions for Web resources of interest to our readers, proposals for columns, as well as general feedback, by e-mail to Dong Yu ("Best of the Web" associate editor) at dongyu@microsoft.com.

In this issue, "Best of the Web" focuses on introducing the social signal processing network (SSPNet), a large European collaboration aimed at establishing a research community in social signal processing (SSP), the new, emerging domain aimed at bringing social intelligence in computers.

One of the most exciting challenges that a researcher can face is to pioneer a new domain and to foster its acceptance and recognition in the scientific community. The success in such an endeavor depends on how promising the new domain is in terms of scientific results but also on a second factor that should not be neglected, namely how difficult it is to enter the new domain for an institution, group, or even an individual researcher. As a matter of fact, entry barriers can prevent even the most interested researchers from entering a fully or largely unexplored domain no matter how interesting and promising the domain is. This is the consideration that drives the efforts of the SSPNet. The SSPNet involves some of the earliest SSP researchers and its ultimate goal is to smooth, if not to eliminate, the three main entry barriers that people face when starting to work on SSP, the lack of knowledge, data, and tools.

The strategy of the SSPNet is reflected on a Web portal (www.sspnet.eu) that aims not only at diffusing information about SSP but also at providing the most important and yet difficult to obtain resources for working in SSP, i.e., the knowledge, data, and tools corresponding to the above-mentioned barriers. The SSPNet portal has been online since August 2009 and, thanks to a collaborative effort involving both SSPNet members (11 institutions scattered across Europe) and contributors from the rest of the scientific community, provides a large amount of SSP-related resources. Both data and tools (see below for more details) are freely available to the scientific community. To share resources through the SSPNet portal is not only an important contribution but also an excellent opportunity for achieving high visibility in the emergent and dynamic community growing around SSP.

In addition to the above-mentioned resources, the portal offers an up-to-date view of the SSP state of the art through an extensive (often updated) bibliography and the Virtual Learning Centre (VLC), a repository of lecture and presentation recordings collected at scientific and training events revolving around SSP. This contributes to the elimination of the last important entry barrier, i.e., the lack of knowledge.

SOCIAL SIGNAL PROCESSING

Social intelligence is the ability of dealing effectively with the complex web of social interactions we cope with in our everyday life and, at its core, it consists of effectively perceiving, correctly understanding, and appropriately reacting to social signals, the complex constellations of nonverbal behavioral cues (e.g., facial expressions, gestures, or vocalizations) through which we express our relational attitudes (e.g., empathy, disagreement, or hostility) with respect to others and social situations.

In this respect, SSP aims to answering the following three main questions:

- Is it possible to detect nonverbal behavioral cues from signals captured through microphones, cameras, or any other suitable sensor?
- Is it possible to automatically infer and understand social signals from nonverbal behavioral cues detected in possibly multimodal signals?
- Is it possible to synthesize appropriate social signals (as a set of synthesized nonverbal behavioral cues) via different forms of embodiment?

In correspondence to the above questions, two main kinds of technologies are involved in SSP: approaches for analysis and synthesis of nonverbal behavioral cues like, e.g., facial expression analysis and synthesis, prosody extraction and synthesis, gesture and posture recognition, and synthesis, as well as techniques for inferring social signals from behavioral cues like, e.g., machine learning and pattern recognition. Equally important for both kinds of technology is the investigation of psychological, anthropological, and social laws underlying human-human interactions. These laws and principles identify the predictable behavioral patterns that actually allow technology to be effective with social signals.

Synthesis and understanding of social signals are mostly data driven, large corpora of data annotated in terms of social signals that become a fundamental resource, hence the data barrier. Furthermore, as nonverbal behavioral cues are typically captured with different sensors (e.g., facial expressions with cameras and vocalizations with microphones), tools addressing a wide

Digital Object Identifier 10.1109/MSP.2010.936780

1053-5888/10/$26.00©2010IEEE

spectrum of diverse needs become crucial, hence the tool barrier. Finally, the multidisciplinary nature of SSP, spanning across multiple technical competences (speech processing and synthesis, computer vision) and human sciences, makes it difficult for a group or even for an institution to have all necessary knowledge at disposition, hence the knowledge barrier.

The three barriers shape the structure of the SSPNet portal and drive the selection of the resources being accumulated in its different sections. The material on the portal is at disposition of the scientific community for research purposes (in some cases upon signing an end user license agreement) and any contribution is welcome as long as it is annotated rigorously (in the case of the data) and relevant to the SSP research. The portal guarantees storage of the material and, most importantly, high visibility in the emergent SSP community.

A slightly modified version of this cartoon appeared in *IEEE Antennas and Propagation Magazine*, vol. 52, no. 1, p. 201, 2010.

RESOURCES TO BREAK THE KNOWLEDGE BARRIER

As SSP is a young domain, its state of the art is still relatively limited, but it is rapidly growing, and it is fragmented across a large number of disciplines and research areas. Thus, it can be difficult for people entering the domain to identify the relevant literature and to access the latest developments in the field. To this end, the SSPNet portal hosts two important sections. The first is an exhaustive bibliography including not only the most important SSP works published so far, but also a large number of works providing the necessary background to enter the field (http://sspnet.eu/category/sspnet_resource_categories/resource_type_classes/publication/). The repository is fully searchable in terms of meta data (title and authors) as well as in terms of tags defined by SSPNet researchers and corresponding to the most important aspects of SSP (http://sspnet.eu/resources/search/) like the behavioral cues being investigated in the article (e.g., "facial analysis" and "speech synthesis"), or the modeling classes ("linguistic modeling" and "psychological modeling"). At the time this column was written, the bibliography included around 300 titles, but it is constantly increasing with contributions from both SSPNet researchers and the rest of the scientific community.

The second important section of the portal dedicated to the knowledge barrier is the VLC (http://sspnet.eu/virtual-learning-centre/), a repository of lecture and presentation recordings collected at scientific events (workshops, special sessions) and training initiatives (summer schools, courses) dedicated to SSP. When this column was being written, the VLC included around 40 presentations collected at the first events organized by the SSPNet. An extensive recording campaign is taking place to further improve depth and breath of the material (the inclusion of 80–100 more presentations is planned by the end of 2010). The VLC is fully searchable in terms of keywords appearing in the presentations slides, a simple textual query (like those submitted to Web search engines) returns presentation intervals corresponding to those slides that are detected as relevant to the query itself.

These two sections of the portal make it possible for any interested researcher to acquire the necessary knowledge about the current state of the art in SSP, including most recent trends, as well as about the background necessary to deal with SSP problems.

RESOURCES TO BREAK THE DATA BARRIER

In SSP, data typically consist of large corpora of video and audio recordings portraying social interactions. The collection of this kind of data is one of the most expensive and time-consuming aspects of SSP. On one hand, recording social interactions often requires large experimental apparatuses like smart meeting rooms, or devices capable of synchronizing multiple sensors. On the other hand, data must be annotated, i.e., trained observers must identify the social phenomena taking place in the recordings, at the exact time when they appear and following a rigorous methodology that allows repeatability of the experiments and a sufficient degree of objectivity (in terms of agreement between multiple annotators). Such a process can take significant amount of time, especially when the corpus is large and the annotation is fine grained (e.g., the annotation of facial expressions requires to track every facial muscle during the time a face is portrayed).

The SSPNet portal hosts a data repository that, at the time this column was being written, contained around 240 h of annotated material (http://sspnet.eu/category/sspnet_resource_categories/resource_type_classes). The data repository includes, among others, the Augmented Multiparty Interaction (AMI) Meeting Corpus (150 meeting recordings annotated in terms of roles, dominance, and subjectivity), the Canal9 Database (75 television debates annotated in terms of conflict, roles, agreement and disagreement), the Belfast Naturalistic Database (298 clips showing 125 speakers in both neutral and emotional states annotated in terms of acoustic features), the Human Communication Research Centre (HCRC) Map Task Corpus (128 task oriented dialogues annotated in terms of discourse phenomena), the IDIAP Head-Pose Database (eight meetings annotated in terms of participant head pose), the Green Persuasive Database (eight dialogues annotated in terms of persuasive behavior), the ICSI Meeting Corpus (75 meeting recordings), the Man-Machine Interaction (MMI) Facial Expression Database (2,894 video clips annotated in terms of facial expressions), and the FreeTalk Corpus (a collection of Japanese phone calls annotated in terms of interactional phenomena).

RESOURCES TO BREAK THE TOOL BARRIER

Once the interested researchers know what SSP is about and have the relevant data at disposition, the last barrier is the lack of suitable tools to perform actual research work. This applies, for instance, to researchers who work on the automatic understanding of social signals but do not have the competences for the extraction of nonverbal behavioral cues or to researchers who know how to process only one modality (e.g., speech) but would like to develop approaches involving other modalities as well (e.g., gestures).

For this reason, the SSPNet portal provides a repository of tools that addresses diverse needs in SSP work (http://sspnet.eu/2009/12/gabor-facial-point-detector/). While this column was being written, the repository contained the Nite XML tool kit (an open source cross-platform framework for handling multimodal annotations that are related both temporally and structurally), a salient point detector for human gestures (it finds spatio-temporal salient points in an image sequence), a real-time gaze and head-pose estimation system (it can use a plain Web camera mounted on top of the user's screen and produces two-dimensional yaw/pitch vectors for the user's eye gaze and head pose, including roll), PRTools (a toolbox for pattern recognition algorithms), the SEMAINE research platform (an open-source software package containing state-of-the-art software tools for audio-visual behavior analysis and synthesis), and the Gabor facial point detector (combining face detection with facial point detection).

ACKNOWLEDGMENTS

This work was funded in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant agreement 231287 (SSPNet). The work of Alessandro Vinciarelli was funded in part by the Swiss National Science Foundation through the National Center of Competence in Research Interactive Multimodal Information Management. The work of Maja Pantic was funded in part by the European Research Council under the ERC Starting Grant agreement ERC-2007-StG-203143 (MANHOB).

AUTHORS

Alessandro Vinciarelli (vincia@dcs.gla.ac.uk) is a lecturer with the Department of Computing Sciences, University of Glasgow, United Kingdom. He is also a senior researcher with Idiap Research Institute, Switzerland.

Maja Pantic (m.pantic@imperial.ac.uk) is a reader in multimodal HCI in the Computing Department with Imperial College London, United Kingdom. She is also a professor of affective and behavioral computing in the Department of Computer Science at the University of Twente, The Netherlands. [SP]

LEARNING HAS NO BOUNDARIES

YOU KNOW YOUR STUDENTS NEED IEEE INFORMATION.
NOW THEY CAN HAVE IT. AND YOU CAN AFFORD IT.

IEEE RECOGNIZES THE SPECIAL NEEDS OF SMALLER COLLEGES, and wants students to have access to the information that will put them on the path to career success. Now, smaller colleges can subscribe to the same IEEE collections that large universities receive, but at a lower price, based on your full-time enrollment and degree programs.

Find out more–visit www.ieee.org/learning

IEEE

[dates AHEAD]

Please send calendar submissions to:
Dates Ahead, c/o Jessica Barragué, *IEEE Signal Processing Magazine* 445 Hoes Lane,
Piscataway, NJ 08855 USA,
e-mail: j.barrague@ieee.org
(Colored conference title indicates SP-sponsored conference.)

2010

[JUNE]

The 11th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2010)
20–23 June, Marrakech, Morocco.
General Cochairs: Mounir Ghogho and Ananthram Swami
URL: http://www.spawc2010.org/

2nd International Workshop on Quality of Multimedia Experience (QoMEX 2010)
21–23 June, Trondheim, Norway.
General Cochairs: Andrew Perkis and Sebastian Möller
URL: http://www.qomex.org/

[JULY]

The IEEE International Conference on Multimedia & Expo (ICME 2010)
19–23 July, Singapore.
General Chairs: Yap-Peng Tan and Oscar C. Au
URL: http://www.icme2010.org

Digital Object Identifier 10.1109/MSP.2010.936722

[AUGUST]

The 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'10)
21–23 August, Beijing, China.
General Chairs: Fuji Ren and Yixin Zhong
URL: http://caai.cn:8080/nlpke10/index.html

The 2010 International Workshop on Machine Learning for Signal Processing (MLSP 2010)
29 August–1 September, Kittilä, Finland.
General Chair: Erkki Oja
URL: http://mlsp2010.conwiz.dk/

[SEPTEMBER]

2010 International Conference on Image Processing (ICIP 2010)
26–29 September, Hong Kong.
General Chair: Wan-Chi Siu
URL: http://www.icip2010.org

[OCTOBER]

The 6th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM '10)
4–7 October, Israel.
General Cochairs: Hagit Messer and Jeffrey L. Krolik
URL: http://www.sam-2010.org/

2010 IEEE Workshop on Signal Processing Systems (SiPS 2010)
6–10 October, San Francisco, California.
General Cochairs: Shuvra Battacharyya and Jorn Janneck
URL: http://www.sips2010.org/

2010 IEEE International Symposium on Phased Array Systems and Technology (ARRAY'10)
12–15 October, Waltham, Massachusetts.
Conference Chair: Mark Russell
URL: http://www.array2010.org/

[NOVEMBER]

2010 2nd International Conference on Audio, Language, and Image Processing
23–25 November 2010, Shanghai, China.
General Chairs: Fa-Long Luo, Wanggen Wan, and Thomas Sikora
URL: http://www.icalip2010.cn/

[DECEMBER]

28th Picture Coding Symposium (PCS'10)
7–10 December, Nagoya, Japan.
URL: http://www.pcs2010.org/

2010 IEEE Spoken Language Technology Workshop (SLT'10)
12–15 December, Berkeley, California.
General Chairs: Dilek Hakkani-Tür and Mari Ostendorf
URL: http://www.slt2010.org/

IEEE **JobSite**

The Right Candidate - Right Now!

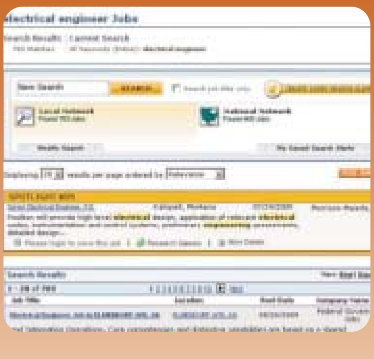
Take the next steps to finding your ideal job.

The IEEE Job Site can help you explore the engineering career opportunities that may be right for you.

- Create a personal, confidential job profile
- Post your skills, desired salary and job location
- Receive e-mail alerts when an employer posts an opening that matches your requirements
- Search job postings at your convenience

Take advantage of this unique member benefit.

Post your resumé on the IEEE Job Site and start your job search now!



Visit the IEEE Job Site at www.ieee.org/jobs

IEEE

continued from 152

controlled device, and rely entirely on the brain to adapt to this fixed mapping. Studies using invasive recording of single neurons or neural populations show that the motor system can rapidly learn to generate appropriate patterns for a fixed mapping (e.g., [3]). However, with EEG in humans, the same approach may take months to achieve an adequate level of performance [4]. Current approaches therefore typically rely on both user adaptation and machine learning. EEG activity patterns are recorded from the user prior to BCI use and this data is utilized to train a pattern recognition algorithm for classification or regression. Data collected from subsequent sessions are then used to further update the classifier or regresser to the user's most recent brain patterns. Simultaneous online adaptation by the user and BCI remains a topic of active research.

SIGNAL TYPES USED IN NONINVASIVE BCIs

The two major types of EEG signals used in BCIs are evoked potentials (EPs) and oscillatory activity patterns. EPs are electrical potential shifts that are phase-locked to external perceptual events such as a rare visual stimulus. EPs are typically analyzed by averaging EEG data over time beginning at the start of the perceptual event for a duration of up to 1 s. Oscillatory activity patterns, on the other hand, can be voluntarily induced by the user, for example, through the imagination of kinaesthetic body movements. Such imagery typically causes a decrease or increase in power in particular frequency bands. This decrease or increase in power is usually referred to as event-related desynchronization (ERD) or event-related synchronization (ERS), respectively.

EEG HAS EMERGED AS THE SINGLE MOST IMPORTANT NONINVASIVE SOURCE OF BRAIN SIGNALS FOR BRAIN-COMPUTER INTERFACING IN HUMANS.

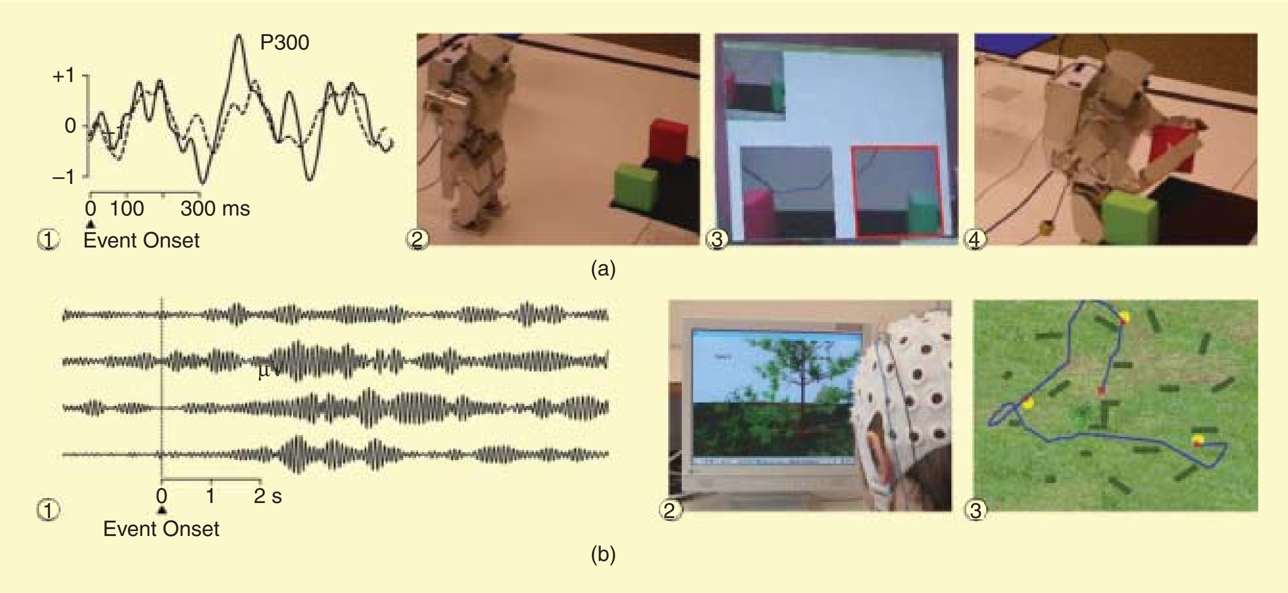
Since EPs are stereotypical brain responses that are stable over time, very little adaptation may be required on the part of the user. Oscillatory patterns of a user, however, typically change over time as a result of feedback during BCI use, making parameters that were learned offline suboptimal. Coadaptation is therefore required: brain signals recorded during feedback are analyzed to track changes in oscillatory patterns and the BCI is updated whenever required.

The two examples below illustrate the use of EPs and oscillatory patterns for achieving brain-computer interaction in physical and virtual environments (VEs).

BRAIN COMPUTER INTERFACING USING EVOKED POTENTIALS

One type of EP that has been used successfully in BCIs is the P300. The P300 is so named because it is characterized by a positive potential shift about 300 ms after the presentation of a perceptually significant event embedded within a series of routine stimuli.

Figure 2(a) illustrates an experimental paradigm that uses the P300 to allow a user to select from a menu of choices. The choices are presented in a grid format on a computer screen. The choices in this experiment correspond to segmented images of objects from the current field of vision of a humanoid



[FIG2] Two examples of noninvasive BCIs. (a) BCI based on EPs. 1) Averaged response over ten trials for attended (P300, solid line) and unattended images (dashed line). 2) Humanoid robot in front of two objects waiting for input from BCI user. 3) User attends to image of desired object while borders are randomly flashed (red square). 4) Humanoid robot picks up the object selected by the user. (b) BCI based on oscillatory activity. 1) Four trials of event related synchronization in the 10–13 Hz frequency band induced by foot motor imagery initiated at time 0 s. 2) BCI user navigating through a VE. 3) Map of the VE showing the trajectory of the BCI user. Yellow markers indicate locations of coins the user was instructed to collect.

helper robot [5]. The P300 (Figure 2(a), Panel 1) is used to infer which object the BCI user would like the robot to pick up for transport to a different location (Figure 2(a), Panels 2–4). Once the object has been picked up, the menu switches to images of possible destination locations and the P300 is again used to infer the user's choice of a destination for the robot.

To make a selection using the P300, the user focuses his or her attention on the image of choice while the borders of the images are flashed one at a time in a random order. Each image is flashed multiple times in this random order. Flashes on the attended image generate P300 responses while the other flashes do not (see Figure 2(a), Panel 1).

A linear support vector machine (SVM) with slack variables [6] was trained to discriminate between P300 and non-P300 responses. Labeled training data for this purpose was obtained in a 10-min data collection protocol at the beginning of the experiment.

The input to the SVM was a low-dimensional feature vector obtained by applying a small set of spatial filters to 32 channels of EEG data recorded from electrodes placed over the entire scalp. These filters are "spatial" because they are applied not to samples over a time period but to the 32 samples spatially distributed over the scalp. The output of a filter is a linear weighted combination of the 32 EEG channels at each time step. Each channel was first band pass filtered in the 0.5–30 Hz range to exclude noise typically present at higher frequencies.

The spatial filters were learned from the labeled data as follows. Let E denote the $32 \times N$ matrix of EEG data, where N is the number of time points (in this case, representing a duration of 500 ms from the onset of each flash). Applying a 32×1 spatial filter f to the 32-channel EEG data results in the following time series of filtered data:

$$x = f^T E.$$

To aid classification, we would like a filter f that maximizes the squared distance between the means of the filtered data for the two classes (P300 and non-P300 responses) while minimizing the within-class variance. This is equivalent to maximizing the criterion

$$J(f) = \frac{tr(S_b)}{tr(S_w)},$$

where tr denotes trace of a matrix, and S_b and S_w are the between-class and within-class scatter matrix, respectively, of the filtered data x. Maximizing J can be shown to be equivalent to a generalized eigenvalue problem whose solution is a set of orthonormal eigenvectors (filters) f ordered by their eigenvalues: the larger the eigenvalue, the more discriminative the filter (see [5] for details). The three filters with the three largest eigenvalues were found to capture most of the discriminative information for the training data. These filters were applied to the 32-channel EEG data to yield three filtered outputs at each time step. This low-dimensional filtered time series data was used to train the classifier.

THE TASK OF THE BCI IS TO IDENTIFY AND PREDICT BEHAVIORALLY INDUCED CHANGES OR "COGNITIVE STATES" IN A USER'S BRAIN SIGNALS.

During the operation of the BCI, the image with the highest number of P300 classifications after the completion of all flashes was selected as the user's choice. An average classification accuracy of 95% across nine users was achieved for discriminating between four choices, using five flashes per choice. With the implemented rate of four flashes per second, the selection of one out of four options takes 5 s, yielding an information transfer rate of 24-b/min.

BCI USING OSCILLATORY ACTIVITY

In the P300-based BCI described above, command generation was synchronized with an externally generated stimulus or cue. Such BCIs are called cue guided. In contrast, BCIs that allow the user to voluntarily modulate brain activity wheneve the user wishes to issue a command are called self-paced. Self-paced BCIs are typically based on detecting changes in oscillatory activity. For example, imagining movements can cause changes in oscillatory EEG activity in the 8–30 Hz frequency range over sensorimotor areas (Figure 2(b), Panel 1). Furthermore, different types of imagined movements can result in different oscillatory patterns which can be classified using machine learning.

As an example, consider navigating in a VE: one could use left hand, right hand, and foot motor imagery to move left, right, and forward, respectively [7]. The subject's task in the experiment was to navigate and find coins that are scattered randomly at different locations in the environment (Figure 2(b), Panels 2–3). A committee of Fisher's linear discriminant analysis (LDA) classifiers [6] with majority voting was trained to discriminate between the three types of motor imagery. An additional LDA classifier was trained to detect whether the subject was engaged in motor imagery or not; only when motor imagery was detected was the committee of classifiers used to predict the type of movement.

Features for classification were estimated from 1-s segments by band-pass filtering the EEG signal for several frequency bands, and squaring and calculating the mean over the squared values for each band in each segment. To decrease variability, features used in classification were based on the logarithm of the band power estimates. The most discriminative frequency bands were identified for each subject independently. To allow real-time interaction, classification was performed every 40 ms. Given the focus on motor imagery, data was recorded from six EEG sensors placed over appropriate sensorimotor areas. Techniques for online muscle artifact detection and eye movement reduction were also used to reduce contamination of the EEG signal (see [7] for details).

After a total of about five hours of coadaptive training over several days, the average three-class accuracy of the LDA

committee classifier reached approximately 80%, with a false positive rate for motor imagery detection (by the additional LDA classifier) of about 17%. Subjects were able to successfully use the BCI to navigate and locate the coins in the environment (Figure 2(b), Panel 3).

NONINVASIVE BCIs: THE FUTURE

The noisy nature of EEG and the fact that brain activity patterns are typically subject-specific means that signal processing and subject-specific optimization are essential for successful brain-computer interaction. The nonstationarity and inherent variability of the EEG, along with limited sample size and limited knowledge about the underlying signal, makes BCIs a challenging domain for signal processing.

Much of past BCI research has focused on cue-based BCIs, where the mental states are more or less well defined. An important challenge for the future is the design and implementation of self-paced BCIs, where a number of distinct patterns have to be reliably detected in ongoing brain activity. Although there have been several prototype systems (e.g., the navigation system discussed above), there is room for improvement.

Another important issue is usability. Current electrode caps and wet electrodes are not practical for everyday use in nonlaboratory settings. Future recording devices will need to be less time consuming to set up, more comfortable to wear, and less expensive to purchase and maintain. The first generation of wireless neuro-signal recording devices with dry electrodes have started appearing on the market (e.g., Emotiv Systems, San Francisco, California). Whether and to what extent these new technologies prove to be useful for BCI applications remains to be seen.

Finally, the problem of coadaptation of brain and machine in BCIs presents an interesting challenge for pattern recognition and machine-learning algorithms. Although some promising preliminary results have been obtained, an overarching theory of coadaptation remains to be developed. Such a theory would entail finding statistical methods that can predict changes in brain activity, allowing the BCI to adapt in sync with the human user for achieving the common goal of direct brain-computer interaction with the external world.

AUTHORS

Rajesh P.N. Rao (rao@cs.washington.edu) is an associate professor in the Department of Computer Science and Engineering at the University of Washington, Seattle.

Reinhold Scherer (scherer@cs.washington.edu) is a posdoctoral research fellow in the Department of Computer Science and Engineering at the University of Washington, Seattle.

REFERENCES

[1] A. Hammock. (2010, Jan. 4). The future of brain-controlled devices. *CNN Online* [Online]. Available: http://www.cnn.com/2009/TECH/12/30/brain.controlled.computers/index.html

[2] G. Dornhege, J. d. R. Millan, T. Hinterberger, D. McFarland, and K. R. Müller, Eds., *Towards Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[3] T. Blakely, K. J. Miller, S. P. Zanos, R. P. N. Rao, and J. G. Ojemann. (2009, July). Robust, long-term control of an electrocorticographic brain-computer interface with fixed parameters. *Neurosurg. Focus* [Online]. 27(1), p. E13. Available: http://thejns.org/doi/full/10.3171/2009.4.FOCUS0977

[4] A. Kübler, B. Kotchoubey, T. Hinterberger, N. Ghanayim, J. Perelmouter, M. Schauer, C. Fritsch, E. Taub, and N. Birbaumer. (1999, Jan.). The thought translation device: A neurophysiological approach to communication in total motor paralysis. *Exp. Brain Res.* [Online]. 124(2), pp. 223–232. Available: http://www.metapress.com/content/FGUE9H81NLPF2JBB

[5] C. J. Bell, P. Shenoy, R. Chalodhorn, and R. P. N. Rao. (2008, June). Control of a humanoid robot by a noninvasive brain–computer interface in humans. *J. Neural Eng.* [Online]. 5(2), pp. 214–220. Available: http://dx.doi.org/10.1088/1741-2560/5/2/012

[6] K. R. Müller, C. W. Anderson, and G. E. Birch. (2003, July). Linear and nonlinear methods for brain-computer interfaces. *IEEE Trans. Neural Syst. Rehab. Eng.* [Online]. 11(2), pp. 165–169. Available: http://dx.doi.org/10.1109/TNSRE.2003.814484

[7] R. Scherer, F. Lee, A. Schlögl, R. Leeb, H. Bischof, and G. Pfurtscheller. (2008). Toward self-paced brain-computer communication: Navigation through virtual worlds. *IEEE Trans. Biomed. Eng.* [Online]. 55(2), pp. 675–682. Available: http://dx.doi.org/10.1109/TBME.2007.903709

[SP]

advertisers INDEX

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*.

COMPANY	PAGE#	URL	PHONE
Aldebaran Robotics	5	www.aldebaran-robotics.com/en/naoresearch	+33 1 7737 1797
DSPE 2011	13	www.dspe2011.org	
Fibertek, Inc.	143	www.opm.gov	+1 703 704 1140
IEEE Marketing	9	www.ieee.org/betterinternet	
IEEE MDL	11	www.ieee.org/go/trymdl	
Mathworks	CVR 4	www.mathworks.com/connect	+1 508 647 7040
Mini-Circuits	CVR 2, 3, CVR 3	www.minicircuits.com	+1 718 934 4500

advertising SALES OFFICES

James A. Vick
Staff Director, Advertising
Phone: +1 212 419 7767;
Fax: +1 212 419 7589
jv.ieeemedia@ieee.org

Marion Delaney
Advertising Sales Director
Phone: +1 415 863 4717;
Fax: +1 415 863 4717
md.ieeemedia@ieee.org

Susan E. Schneiderman
Business Development Manager
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org

Product Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, PA, DE, MD, DC, KY, WV

NEW ENGLAND/ EASTERN CANADA
Jody Estabrook
Phone: +1 774 283 4528;
Fax: +1 774 283 4527
je.ieeemedia@ieee.org
ME, VT, NH, MA, RI, CT
Canada: Quebec, Nova Scotia,
Newfoundland, Prince Edward Island,
New Brunswick

SOUTHEAST
Thomas Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
tf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/CENTRAL CANADA
Dave Jones
Phone: +1 708 442 5633;
Fax: +1 708 442 7620
dj.ieeemedia@ieee.org
IL, IA, KS, MN, MO, NE, ND,
SD, WI, OH
Canada: Manitoba,
Saskatchewan, Alberta

MIDWEST/ ONTARIO, CANADA
Will Hamilton
Phone: +1 269 381 2156;
Fax: +1 269 381 2556
wh.ieeemedia@ieee.org
IN, MI. Canada: Ontario

SOUTHWEST
Shaun Mehr
Phone: +1 949 923 1660;
Fax: +1 775 908 2104
sm.ieeemedia@ieee.org
AR, LA, OK, TX

WEST COAST/ NORTHWEST/ WESTERN CANADA
Marshall Rubin
Phone: +1 818 888 2407;
Fax: +1 818 888 4907
mr.ieeemedia@ieee.org
AZ, CO, HI, NM, NV, UT, AK, ID, MT,
WY, OR, WA, CA. Canada: British
Columbia

EUROPE/AFRICA/MIDDLE EAST
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

ASIA/FAR EAST/PACIFIC RIM
Susan Schneiderman
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org
Asia, Far East, Pacific Rim,
Australia, New Zealand

Recruitment Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, CT, PA, DE, MD, DC, KY, WV

NEW ENGLAND/EASTERN CANADA
John Restchack
Phone: +1 212 419 7578;
Fax: +1 212 419 7589
j.restchack@ieee.org
ME, VT, NH, MA, RI. Canada: Quebec,
Nova Scotia, Prince Edward Island,
Newfoundland, New Brunswick

SOUTHEAST
Cathy Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
cf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/TEXAS/CENTRAL CANADA
Darcy Giovingo
Phone: +1 847 498 4520;
Fax: +1 847 498 5911
dg.ieeemedia@ieee.org;
AR, IL, IN, IA, KS, LA, MI, MN, MO, NE,
ND, SD, OH, OK, TX, WI. Canada:
Ontario, Manitoba, Saskatchewan, Alberta

WEST COAST/SOUTHWEST/ MOUNTAIN STATES/ASIA
Tim Matteson
Phone: +1 310 836 4064;
Fax: +1 310 836 4067
tm.ieeemedia@ieee.org
AZ, CO, HI, NV, NM, UT, CA, AK, ID, MT,
WY, OR, WA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

Digital Object Identifier 10.1109/MSP.2010.936721

Rajesh P.N. Rao and
Reinhold Scherer

Brain-Computer Interfacing

Recently, CNN reported on the future of brain-computer interfaces (BCIs) [1]. BCIs are devices that process a user's brain signals to allow direct communication and interaction with the environment. BCIs bypass the normal neuromuscular output pathways and rely on digital signal processing and machine learning to translate brain signals to action (Figure 1). Historically, BCIs were developed with biomedical applications in mind, such as restoring communication in completely paralyzed individuals and replacing lost motor function. More recent applications have targeted nondisabled individuals by exploring the use of BCIs as a novel input device for entertainment and gaming.

The task of the BCI is to identify and predict behaviorally induced changes or "cognitive states" in a user's brain signals. Brain signals are recorded either noninvasively from electrodes placed on the scalp [electroencephalogram (EEG)] or invasively from electrodes placed on the surface of or inside the brain. BCIs based on these recording techniques have allowed healthy and disabled individuals to control a variety of devices [2]. In this article, we will describe different challenges and proposed solutions for noninvasive brain-computer interfacing.

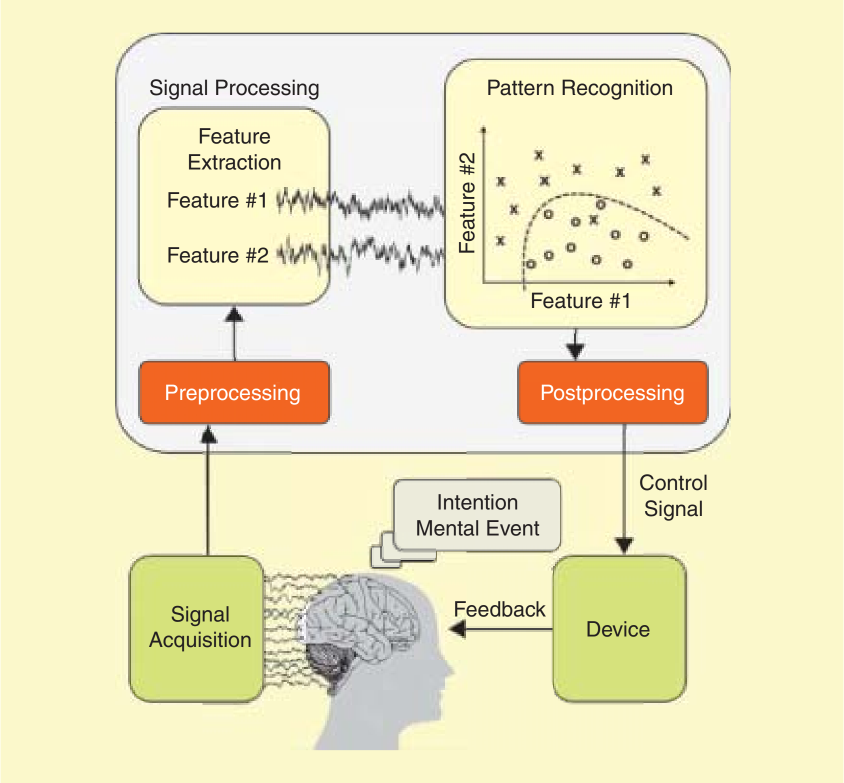
CHALLENGES IN NONINVASIVE BRAIN COMPUTER INTERFACING

EEG has emerged as the single most important noninvasive source of brain signals for brain-computer interfacing in humans. Two major problems confronting BCI developers using EEG are its nonstationarity and its inherent variability. Data from the same experimental paradigm but recorded on different days or even different times on the same day are likely to exhibit differences due to, for instance, shifts in electrode positions between sessions or changes in electromechanical properties of the electrodes (e.g., changing impedances). Additionally, the noisy, nonlinear superposition of the electrical activity of large populations of neurons as measured on the scalp can mask the underlying neural patterns and hamper their detection. The user's current mental state (e.g., due to excessive workload or stress) may impact the ability to focus and generate specific mental events. Due to these factors, statistical signal processing and machine learning techniques play a crucial role in recognizing EEG patterns and translating them into control signals.

BRAIN-COMPUTER COADAPTATION

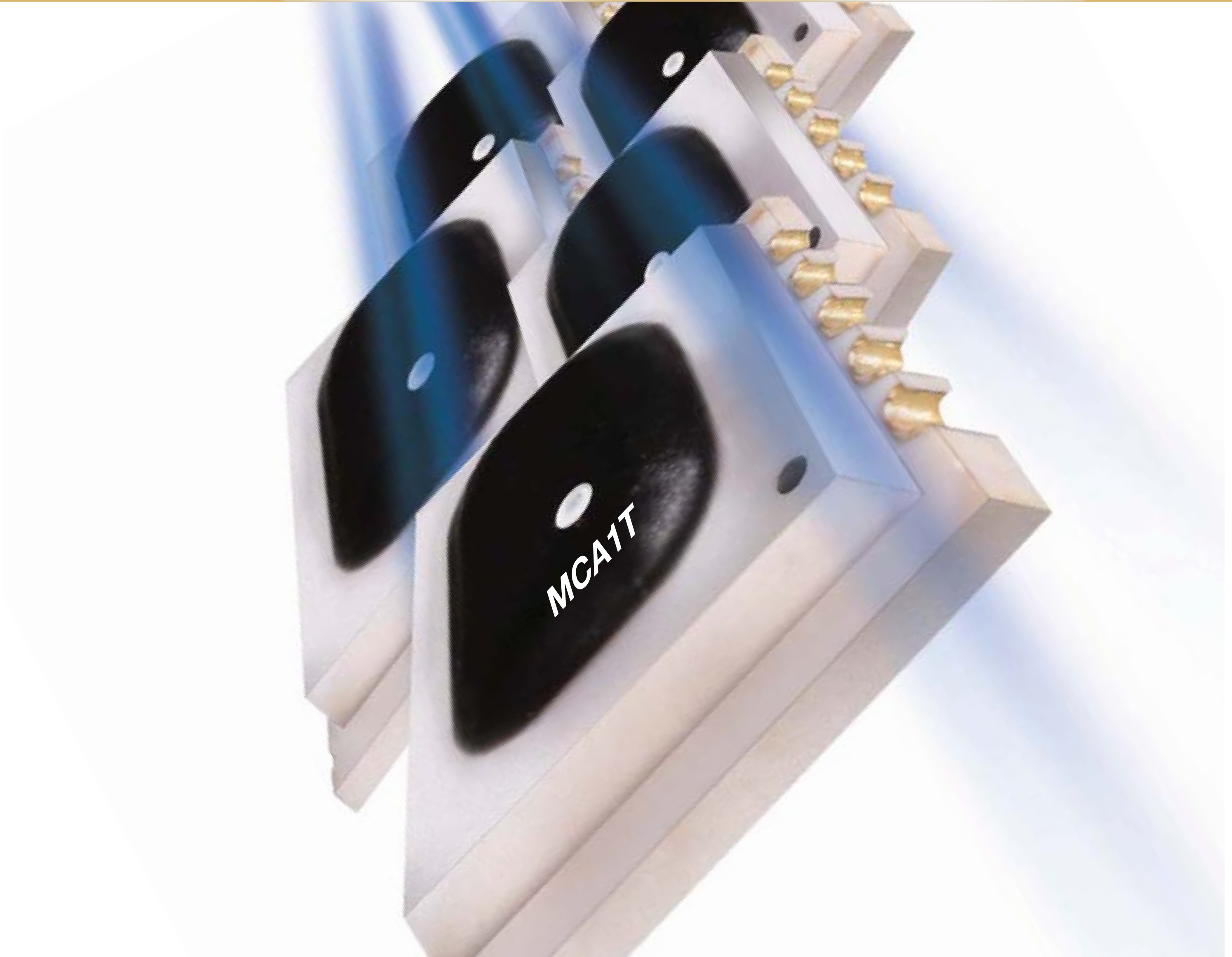
An interesting problem confronting BCI developers is that the brain itself is a highly adaptive device, raising the question of how much of the learning should be relegated to the machine and how much should be left to the brain. At one extreme are approaches that fix a mapping a priori between the brain and a

(continued on page 148)



[FIG1] Basic components of a BCI. Brain activity is translated into a control signal for an external device using a sequence of processing stages. The user receives feedback from the device, thereby closing the loop.

Digital Object Identifier 10.1109/MSP.2010.936774



Ceramic Mixers

300 MHz-12 GHz

- *New gold connections*
- *Ceramic package gives superior reliability*
- *High performance at a lower price*

from $2.95 *ea. (qty. 1000)*

The patented **MCA1T** series of ceramic mixers now feature castellated gold-over-nickel plate connections. These drop-in replacements have the same footprint as our leaded models and offer the same high level of wideband performance with an even higher level of reliability.

Pick from 16 models with LO levels from 4 to 17 dBm and isolation up to 40 dB. **MCA1T** mixers contain their circuitry in a compact, tough-as-nails ceramic package, making them temperature-stable and able to withstand most environmental conditions. And the variety of wideband models means you can use these mixers in many different applications.

Detailed technical specifications are available at minicircuits.com. And, as always, Mini-Circuits is ready to help with quick, off-the-shelf shipments, fast turnaround, and even custom design.

Mini-Circuits...Your partners for success since 1969

	LO	Freq.	Conv.	LO-RF	Price
	Level	Range	Loss	Isol.	$ ea.
MODEL	(dBm)	(MHz)	(dB)	(dB)	(Qty.10-49)
MCA1T-85L+	4	2800-8500	6.0	35	7.95
MCA1T-12GL+	4	3800-12000	6.8	38	10.45
MCA1T-24+	7	300-2400	6.1	40	4.95
MCA1T-42+	7	1000-4200	6.1	35	5.95
MCA1T-60+	7	1600-6000	6.2	32	6.45
MCA1T-85+	7	2800-8500	5.6	37	7.45
MCA1T-12G+	7	3800-12000	6.2	38	9.45
MCA1T-24LH+	10	300-2400	6.5	40	5.45
MCA1T-42LH+	10	1000-4200	6.0	38	5.95
MCA1T-60LH+	10	1700-6000	6.6	35	6.95
MCA1T-80LH+	10	2800-8000	6.0	35	8.45
MCA1T-24MH+	13	300-2400	6.1	40	5.95
MCA1T-42MH+	13	1000-4200	6.2	35	6.45
MCA1T-60MH+	13	1700-6000	6.4	27	7.45
MCA1T-80MH+	13	2800-8000	5.7	27	9.45
MCA1T-80H+	17	2800-8000	6.3	35	10.45

Dimensions: (L) 0.35" x (W) 0.28" x (H) 0.09"

U.S. Patent # 7,027,795

Mini-Circuits®

ISO 9001 ISO 14001 AS 9100 CERTIFIED

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

YONI2™ Patent Pending **The Design Engineers Search Engine** *finds the model you need, Instantly* • For detailed performance specs & shopping online see minicircuits.com

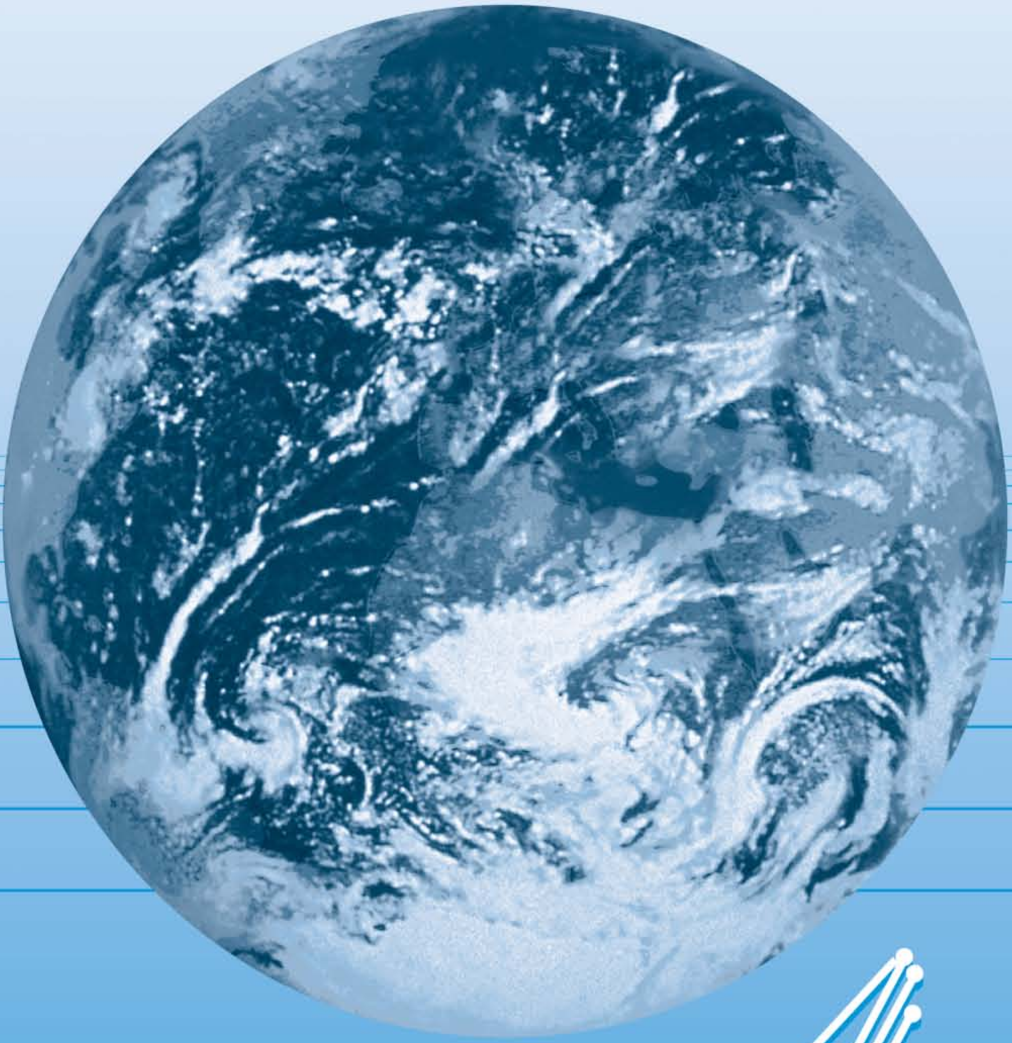
IF/RF MICROWAVE COMPONENTS

479 Rev Orig.

IEEE SIGNAL PROCESSING SOCIETY

ContentGazette

JULY 2010



IEEE Signal Processing Society ®

IEEE

IEEE Signal Processing Society
14th DSP Workshop & 6th SPE Workshop
Enchantment Resort , Sedona, Arizona
January 4-7, 2011
www.dspe2011.org

Organizing Committee

General Chairs
Lina Karam, Arizona State University
Ronald Schafer, Hewlett-Packard Labs

DSP Technical Program Chairs
James McClellan, Georgia Tech
Ali Sayed, UCLA

SPE Technical Program Chairs
Gail Rosen, Drexel University
Thad Welch, Boise State University

Advisory Committee
Ahsan Aziz, National Instruments
Khaled El-Maleh, Qualcomm
Gene Frantz, Texas Instruments
Loren Shure, Mathworks
Mark J.T. Smith, Purdue
Martin Vetterli, EPFL

Finance
David Frakes, Arizona State University

Publicity
Andreas Spanias, Arizona State University

Social Programs
Cathy Wicks, Texas Instruments

International Liaisons
Julien Epps, UNSW, Australia
Ramón Rodríguez Dagnino, ITESM, Mexico
Hideaki Sakai, Kyoto University, Japan
Abdelhak Zoubir, Darmstadt Univ., Germany

IEEE

IEEE Signal Processing Society

Call for Papers

The 2011 IEEE Digital Signal Processing (DSP) Workshop and IEEE Signal Processing Education (SPE) Workshop will be held jointly January 4 to 7, 2011, at the award-winning Enchantment Resort. The Enchantment Resort is located 5 miles from Sedona in Boynton Canyon, two hours north of the Phoenix / Scottsdale metropolitan area and two and a half hours south of the Grand Canyon. The venue is surrounded by the Coconino National Forest and Red Rock Secret Mountain Wilderness. The area is revered by the Apache Native Americans as the birthplace of their tribe and holds ancient ruins of Native American cliff dwellings.

The goals of the workshops are to bring together leading engineers, researchers, and educators in signal processing from around the world to discuss novel signal processing theories, methods, and applications. The DSP/SPE Workshops will feature prominent plenary speakers from the signal processing community as well as technical sessions for presenting contributed papers.

Topics for the DSP Workshop include, but are not limited to:

- Sampling, extrapolation, and interpolation
- System modeling, representation, and identification; deconvolution
- Filtering and adaptive systems
- Stationary signals and spectral analysis
- Non-stationary signals and time-frequency analysis
- Multi-rate signal processing and wavelets
- Detection, estimation, and classification
- Signal enhancement, restoration, and reconstruction
- Nonlinear signal processing
- Multi-dimensional signal processing; image and video processing
- Implementations of Signal Processing Systems
- Distributed signal processing
- New directions and applications

Topics for the SPE Workshop include, but are not limited to:

- Signal processing education in non-traditional venues
- Novel laboratory, computer-based, and distance teaching methods
- Signal processing across the engineering curriculum
- DSP curriculum issues (early/late, simulation/real-time, theory/practice)
- DSP outreach issues

Paper Submission: Prospective authors are invited to submit double-column papers of no more than six (6) pages including title, authors' names and contact, abstract, introduction, background, proposed method, results, figures, and references. Submission instructions and templates for the required paper format are available at www.dspe2011.org.

Important Deadlines:
Submission of Papers: August 30, 2010
Notification of Acceptance: October 11, 2010
Authors' Registration Deadline: October 25, 2010
Submission of Accepted Camera-Ready Papers: November 8, 2010.
Advance Registration and Resort Reservation Deadline: November 15, 2010

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

JUNE 2010 VOLUME 58 NUMBER 6 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

Statistical Signal Processing

Adaptive Signal Processing

Digital and Multirate Signal Processing

Machine Learning

Multimedia Signal Processing

SAM Sensor Array and Multichannel Processing

Signal Processing for Communications

MIMO Communications & Signal Processing

Signal Processing for Sensor Networks

Biomedical Signal Processing

Implementation of Signal Processing Systems

Other Areas and Applications

CORRESPONDENCE

Statistical Signal Processing

Digital and Multirate Signal Processing

Signal Processing for Communications

MIMO Communications & Signal Processing

Signal Processing for Sensor Networks

IEEE Workshop on Spoken Language Technology

SLT 2010

December 12-15, 2010
Berkeley, CA

www.slt2010.org

IEEE

IEEE Signal Processing Society

Organizing Chairs:
DilekHakkani-Tür, ICSI
Mari Ostendorf, U. Washington
Finance Chair:
GokhanTur, SRI International
Advisory Board:
Mazin Gilbert, AT&T Labs - Research
Srinivas Bangalore, AT&T Labs - Research
Giuseppe Riccardi, U. Trento
Technical Chairs:
Isabel Trancoso, INESC-ID, Portugal
Tim Paek, Microsoft
Demo Chairs:
Alex Potamianos, Tech. U. of Crete
MikkoKurimo, Helsinki U. of Tech.
Publicity Chair:
BhuvanaRamabhadran, IBM
Panel Chairs:
SadaokiFurui, Tokyo Inst. Of Tech.
Eric Fosler-Lussier, Ohio State U.
Publication Chair:
Yang Liu, U. Texas, Dallas
Local Organizers:
DimitraVergryi, SRI International
Murat Akbacak, SRI International
SibelYaman, ICSI
Benoit Favre, ICSI
Europe Liaisons:
Frederic Bechet, U. Avignon
Philipp Koehn, U. Edinburgh
Asia Liaisons:
Helen Meng, C. U. Hong Kong
Gary Geunbae Lee, POSTECH

Call for Papers

The Third IEEE Spoken Language Technology (SLT) workshop will be held from December 12 to December 15, 2010 in Berkeley, CA. The goal of this workshop is to allow the spoken language processing community to share and present recent advances in various areas of spoken language technology.

Workshop Topics

- Spoken language understanding
- Spoken document summarization
- Machine translation for speech
- Spoken language based systems
- Spoken language generation
- Question answering from speech
- Human/Computer Interaction
- Speech data mining
- Spoken information extraction
- Spoken document retrieval
- Multimodal processing
- Spoken dialog systems
- Spoken language systems
- Spoken language databases

Submissions for the Technical Program

The workshop program will consist of tutorials, oral and poster presentations, and panel discussions. Prospective authors are invited to submit full-length to the SLT 2010 website http://www.slt2010.org

All papers will be handled and reviewed electronically. The website will provide you with further details.

Important Dates	
Paper submission	July 16, 2010
Paper acceptance/rejection	September 1, 2010
Workshop	December 12-15, 2010

Call for Papers
IEEE Transactions on Audio, Speech, and Language Processing
Special Issue on Deep Learning for Speech Recognition and Related Applications

Over the past 25 years or so, speech recognition technology has been dominated largely by hidden Markov models (HMMs). Significant technological success has been achieved using complex and carefully engineered variants of HMMs. Next generation technologies require solutions to technical challenges presented by diversified deployment environments. These challenges arise from the many types of variability present in the speech signal itself. Overcoming these challenges is likely to require "deep" architectures with efficient and effective learning algorithms.

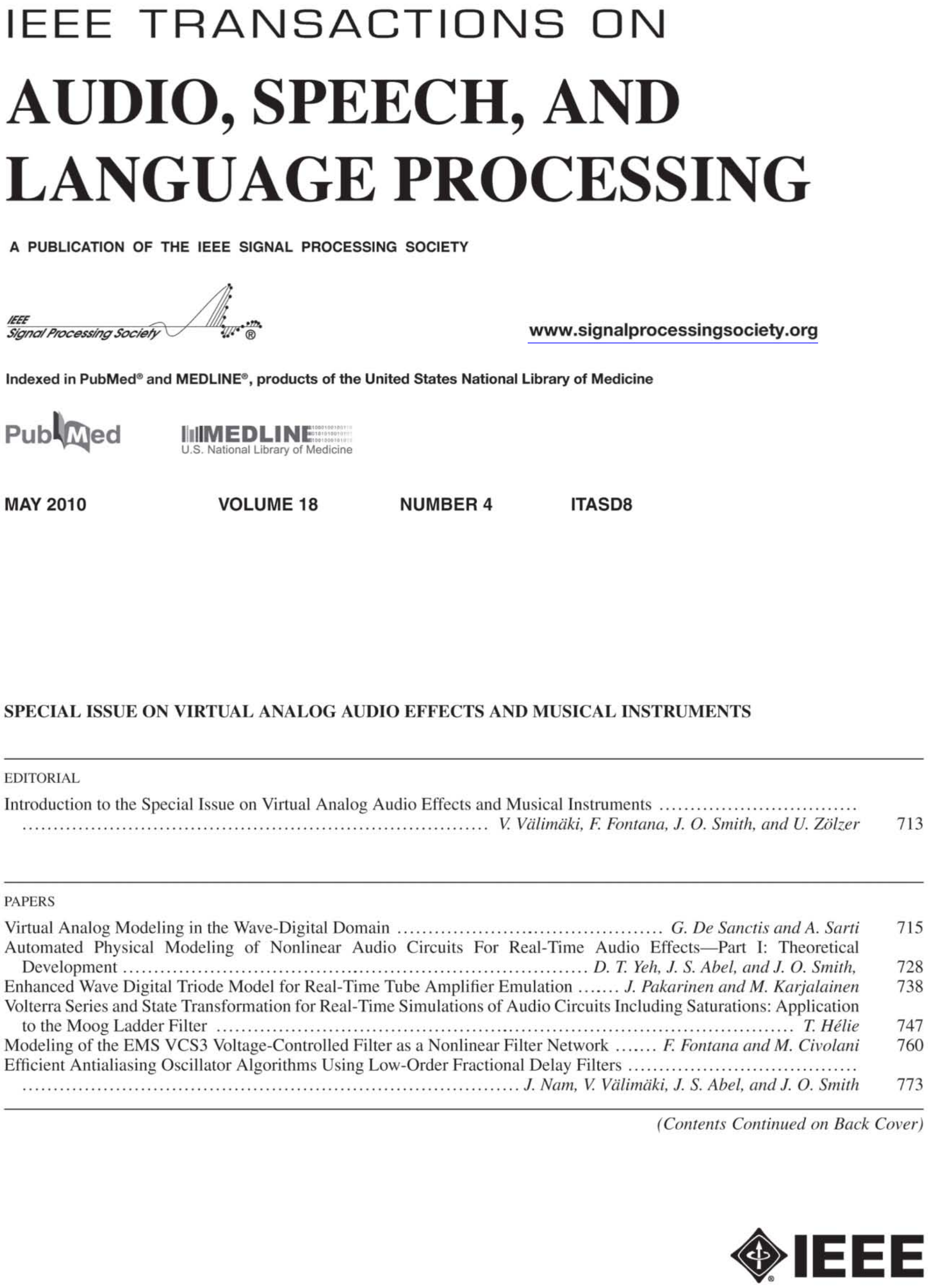
There are three main characteristics in the deep learning paradigm: 1) layered architecture; 2) generative modeling at the lower layer(s); and 3) unsupervised learning at the lower layer(s) in general. For speech recognition and related sequential pattern recognition applications, some attempts have been made in the past to develop layered computational architectures that are "deeper" than conventional HMMs, such as hierarchical HMMs, hierarchical point-process models, hidden dynamic models, layered multilayer perceptron, tandem-architecture neural-net feature extraction, multi-level detection-based architectures, deep belief networks, hierarchical conditional random field, and deep-structured conditional random field. While positive recognition results have been reported, there has been a conspicuous lack of systematic learning techniques and theoretical guidance to facilitate the development of these deep architectures. Further, there has been virtually no effective communication between machine learning researchers and speech recognition researchers who are both advocating the use of deep architecture and learning. The recent NIPS Workshop (Dec 12, 2009; http://research.microsoft.com/en-us/um/people/dongyu/NIPS2009/) successfully brought together these two groups of researchers to review the progress in both fields. The participants of the workshop also presented a wealth of research results pertaining to insightful applications of deep learning to some classical speech recognition and language processing problems. They further identified a set of promising and synergistic research directions for potential future cross-fertilization and collaboration so as to advance the state of the arts in speech and language processing.

In light of the sufficient research activities in this exciting space already taken place and their importance, we invite papers describing various aspects of deep learning and related techniques/architectures as well as their successful applications to speech and language processing. Submissions must not have been previously published, with the exception that substantial extensions of conference or workshop papers will be considered.

This is an open call for papers, inviting prospective authors outside of the Workshop as well as those participating and contributing to the Workshop. The prospective authors can find submission information at http://research.microsoft.com/TBD/. The authors are required to follow the Author's Guide for manuscript submission to the IEEE Transactions on Audio, Speech, and Signal Processing at http://ewh.ieee.org/soc/sps/TBD/.

Submission deadline: **August 27, 2010**
First round of reviews completed: November 26, 2010
Revised manuscripts due: December 30, 2010
Second round of reviews completed: January 28, 2011
Final manuscripts due: March 31, 2011

Lead Guest editor:	**Dong Yu** (Email: dongyu@microsoft.com)
Guest Editors:	**Li Deng** (Email: deng@microsoft.com)
	Geoffrey Hinton (Email: hinton@cs.toronto.edu**)**
	Nelson Morgan (E-mail: morgan@ICSI.Berkeley.edu)
	Jen-Tzung Chien (E-mail: jtchien@mail.ncku.edu.tw)



IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE 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

MAY 2010 VOLUME 18 NUMBER 4 ITASD8

SPECIAL ISSUE ON VIRTUAL ANALOG AUDIO EFFECTS AND MUSICAL INSTRUMENTS

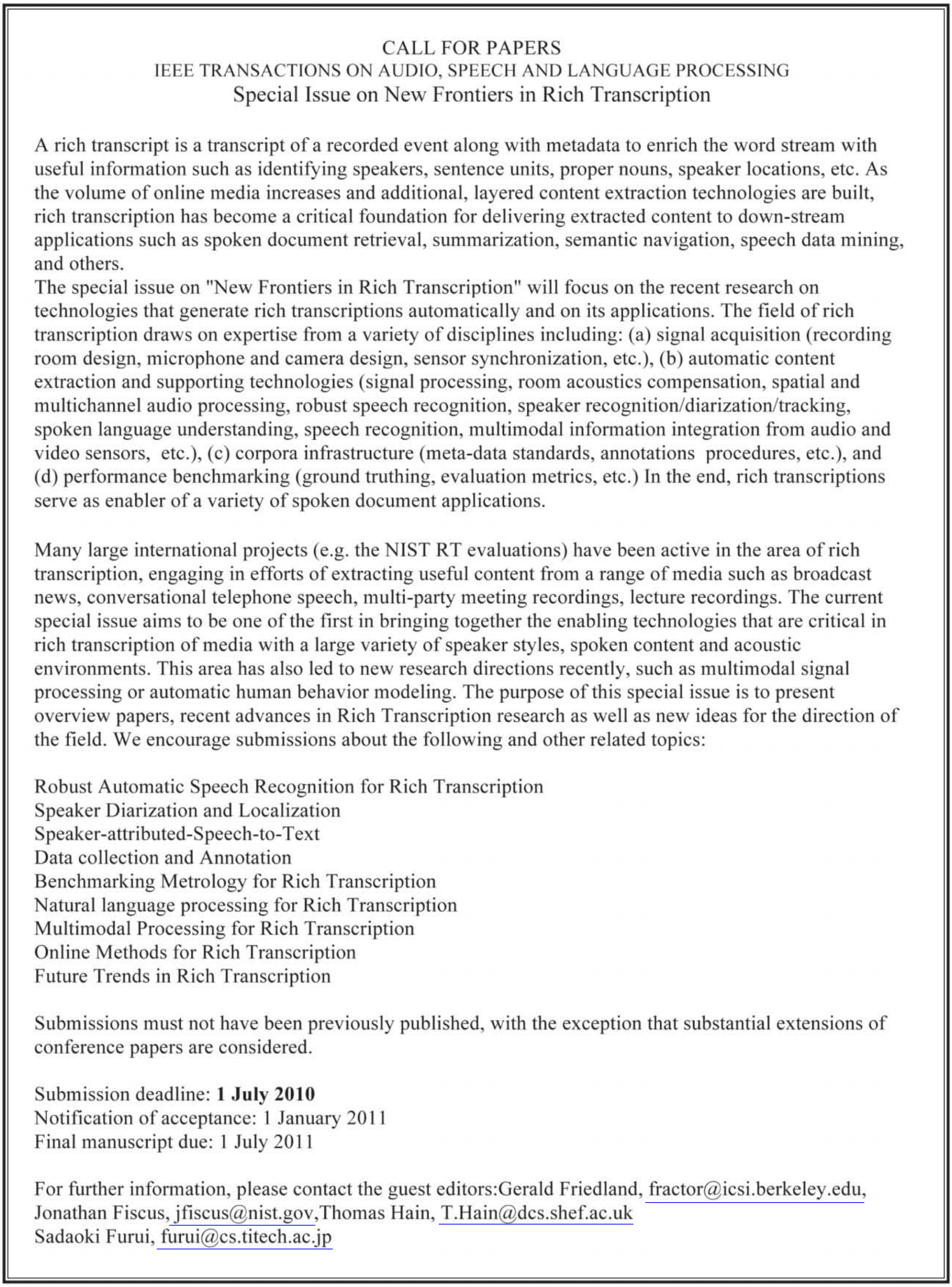
EDITORIAL

PAPERS

(*Contents Continued on Back Cover*)

IEEE

(Contents Continued from Front Cover)



CALL FOR PAPERS
IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING
Special Issue on New Frontiers in Rich Transcription

A rich transcript is a transcript of a recorded event along with metadata to enrich the word stream with useful information such as identifying speakers, sentence units, proper nouns, speaker locations, etc. As the volume of online media increases and additional, layered content extraction technologies are built, rich transcription has become a critical foundation for delivering extracted content to down-stream applications such as spoken document retrieval, summarization, semantic navigation, speech data mining, and others.

The special issue on "New Frontiers in Rich Transcription" will focus on the recent research on technologies that generate rich transcriptions automatically and on its applications. The field of rich transcription draws on expertise from a variety of disciplines including: (a) signal acquisition (recording room design, microphone and camera design, sensor synchronization, etc.), (b) automatic content extraction and supporting technologies (signal processing, room acoustics compensation, spatial and multichannel audio processing, robust speech recognition, speaker recognition/diarization/tracking, spoken language understanding, speech recognition, multimodal information integration from audio and video sensors, etc.), (c) corpora infrastructure (meta-data standards, annotations procedures, etc.), and (d) performance benchmarking (ground truthing, evaluation metrics, etc.) In the end, rich transcriptions serve as enabler of a variety of spoken document applications.

Many large international projects (e.g. the NIST RT evaluations) have been active in the area of rich transcription, engaging in efforts of extracting useful content from a range of media such as broadcast news, conversational telephone speech, multi-party meeting recordings, lecture recordings. The current special issue aims to be one of the first in bringing together the enabling technologies that are critical in rich transcription of media with a large variety of speaker styles, spoken content and acoustic environments. This area has also led to new research directions recently, such as multimodal signal processing or automatic human behavior modeling. The purpose of this special issue is to present overview papers, recent advances in Rich Transcription research as well as new ideas for the direction of the field. We encourage submissions about the following and other related topics:

Robust Automatic Speech Recognition for Rich Transcription
Speaker Diarization and Localization
Speaker-attributed-Speech-to-Text
Data collection and Annotation
Benchmarking Metrology for Rich Transcription
Natural language processing for Rich Transcription
Multimodal Processing for Rich Transcription
Online Methods for Rich Transcription
Future Trends in Rich Transcription

Submissions must not have been previously published, with the exception that substantial extensions of conference papers are considered.

Submission deadline: **1 July 2010**
Notification of acceptance: 1 January 2011
Final manuscript due: 1 July 2011

For further information, please contact the guest editors:Gerald Friedland, fractor@icsi.berkeley.edu,
Jonathan Fiscus, jfiscus@nist.gov,Thomas Hain, T.Hain@dcs.shef.ac.uk
Sadaoki Furui, furui@cs.titech.ac.jp

ICIP 2010

2010 International Conference on Image Processing (ICIP)

26-29 September 2010

The Hong Kong Convention and Exhibition Centre, Hong Kong

Leave room on *YOUR* calendars registration starts

Soon.......

Visit www.icip2010.org for more information

Stay tuned for the announcement of the
2016 ICIP Call for Proposal in the next issue of the Content
Gazette

IEEE

IEEE Signal Processing Society

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 2010 VOLUME 19 NUMBER 6 IIPRE4 (ISSN 1057-7149)

PAPERS

Statistical-Model Based Methods

Perception and Quality Models for Images and Video

Multiresolution Processing of Images and Video

Restoration and Enhancement

Interpolation, Super-Resolution, and Mosaicing

Error Resilience and Channel Coding for Image and Video Systems

IEEE

Image and Video Processing for Watermarking and Security

Color and Multispectral Imaging

Stereoscopic and Multiview Processing and Display

Magnetic Resonance Imaging

Region, Boundary, and Shape Analysis

Image and Video Mid Level Analysis

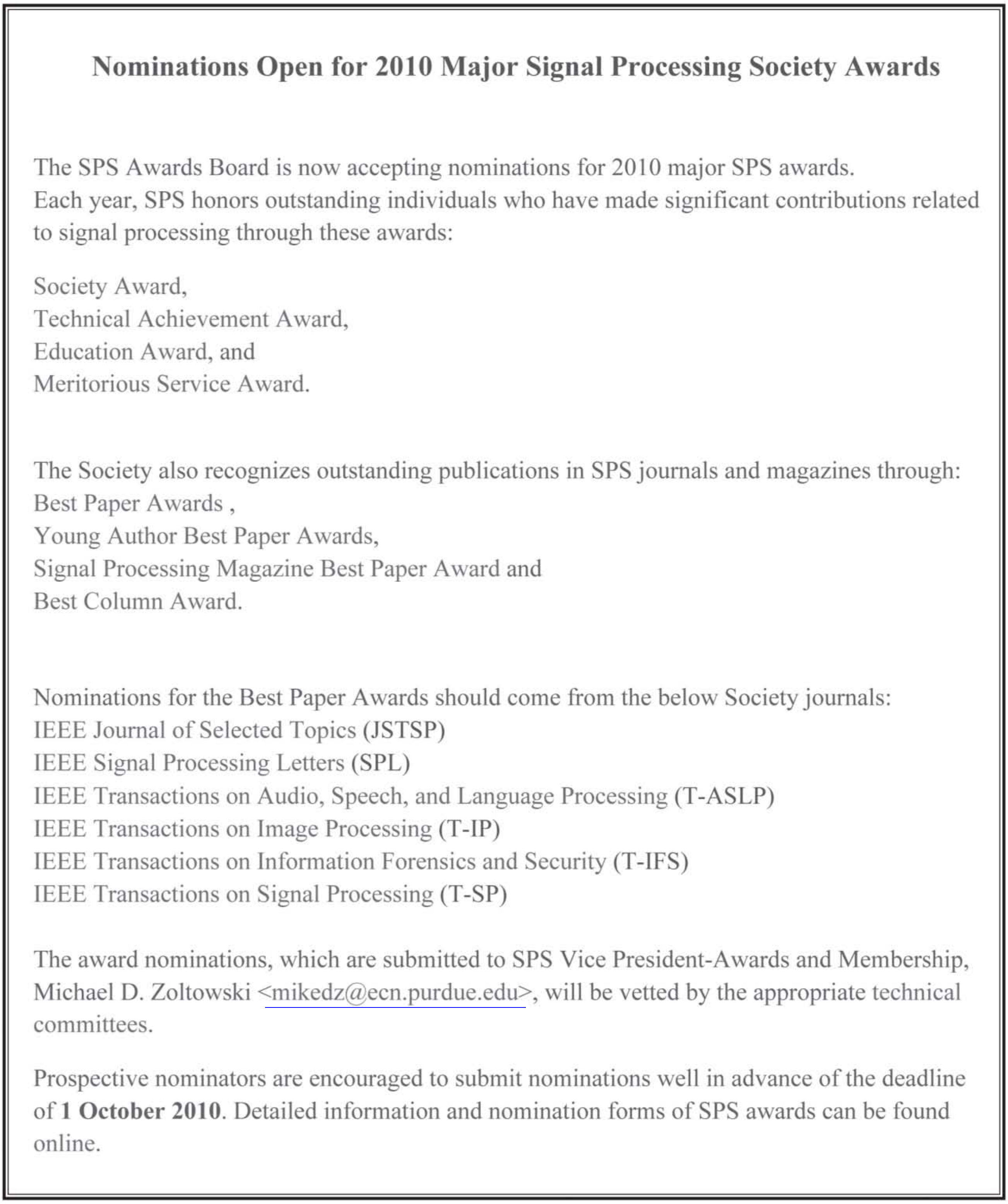
Image and Video Biometric Analysis

CORRESPONDENCE

Lossy Coding of Images and Video

Image and Video Storage and Retrieval

ANNOUNCEMENTS



Nominations Open for 2010 Major Signal Processing Society Awards

The SPS Awards Board is now accepting nominations for 2010 major SPS awards.
Each year, SPS honors outstanding individuals who have made significant contributions related to signal processing through these awards:

Society Award,
Technical Achievement Award,
Education Award, and
Meritorious Service Award.

The Society also recognizes outstanding publications in SPS journals and magazines through:
Best Paper Awards ,
Young Author Best Paper Awards,
Signal Processing Magazine Best Paper Award and
Best Column Award.

Nominations for the Best Paper Awards should come from the below Society journals:
IEEE Journal of Selected Topics (JSTSP)
IEEE Signal Processing Letters (SPL)
IEEE 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)

The award nominations, which are submitted to SPS Vice President-Awards and Membership, Michael D. Zoltowski <mikedz@ecn.purdue.edu>, will be vetted by the appropriate technical committees.

Prospective nominators are encouraged to submit nominations well in advance of the deadline of **1 October 2010**. Detailed information and nomination forms of SPS awards can be found online.

Call for Papers
IEEE Transactions on Information Forensics and Security
Special Issue on Using the Physical Layer for Securing the Next Generation of Communication Systems

Communication technologies are undergoing a renaissance as there is a movement to explore new, clean slate approaches for building communication networks. Although future Internet efforts promise to bring new perspectives on protocol designs for high-bandwidth, access-anything from anywhere services, ensuring that these new communication systems are secure will also require a re-examination of how we build secure communication infrastructures. Traditional approaches to building and securing networks are tied tightly to the concept of protocol layer separation. For network design, routing is typically considered separately from link layer functions, which are considered independently of transport layer phenomena or even the applications that utilize such functions. Similarly, in the security arena, MAC-layer security solutions (e.g. WPA2 for 802.11 devices) are typically considered as point-solutions to address threats facing the link layer, while routing and transport layer security issues are dealt with in distinct, non-integrated protocols like IPSEC and TLS. The inherent protocol separation involved in security solutions is only further highlighted by the fact that the physical layer is generally absent from consideration.

This special issue seeks to provide a venue for ongoing research area in physical layer security across all variety of communication media, ranging from wireless networks at the edge to optical backbones at the core of the network. The scope of this special issue will be interdisciplinary, involving contributions from experts in the areas of cryptography, computer security, information theory, signal processing, communications theory, and propagation theory. In particular, the areas of interest include, but are not limited to, the following:

- Information-theoretic formulations for confidentiality and authentication
- Generalizations of Wyner's wiretap problem to wireless and optical systems
- Physical layer techniques for disseminating information
- Techniques to extract secret keys from channel state information
- Secrecy of MIMO and multiple-access channels
- Physical layer methods for detecting and thwarting spoofing and Sybil attacks
- Techniques to achieve covert or stealthy communication at the physical layer
- Quantum cryptography
- Modulation recognition and forensics
- Security and trustworthiness in cooperative communication
- Fast encryption using physical layer properties
- Attacks and threat analyses targeted at subverting physical layer communications

Manuscript Submission: Manuscripts are to be submitted according to the Information for Authors at http://www.signalprocessingsociety.org/publications/periodicals/forensics/forensics-authors-info/, using the IEEE online manuscript system, Manuscript Central. Papers must not have appeared elsewhere, and must not be in review elsewhere. All papers will be reviewed in accordance with the procedures of the IEEE Transactions. If necessary, the submission date can be moved later based on when the proposal is approved.

Submission deadline: **September 15, 2010**
First Review: December 1, 2010
Revisions Due: January 30, 2011
Final Decision: February 15, 2011
Final manuscript due: March 1,2011
Tentative publication date: June 1, 2011

Guest Editors:
Vincent Poor, Princeton University, (Email: poor@princeton.edu)
Wade Trappe, WINLAB, Rutgers University, (Email: trappe@winlab.rutgers.edu)
Aylin Yener, Pennsylvania State University, (Email: yener@engr.psu.edu)
Hisato Iwai, Doshisha University, Japan, (Email: iwai@mail.doshisha.ac.jp)
Joao Barros, University of Porto, Portugal, (Emal: jbarros@fe.up.pt)
Paul Prucnal, Princeton University, (Email: prucnal@princeton.edu)

IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

JUNE 2010 VOLUME 5 NUMBER 2 ITIFA6 (ISSN 1556-6013)

PAPERS

Watermarking and Data Hiding/Embedding

Cryptographic and Related Techniques

Biometrics

Forensic Analysis

Security and Privacy Analysis

IEEE

CORRESPONDENCE

Watermarking and Data Hiding/Embedding

ANNOUNCEMENTS

IEEE

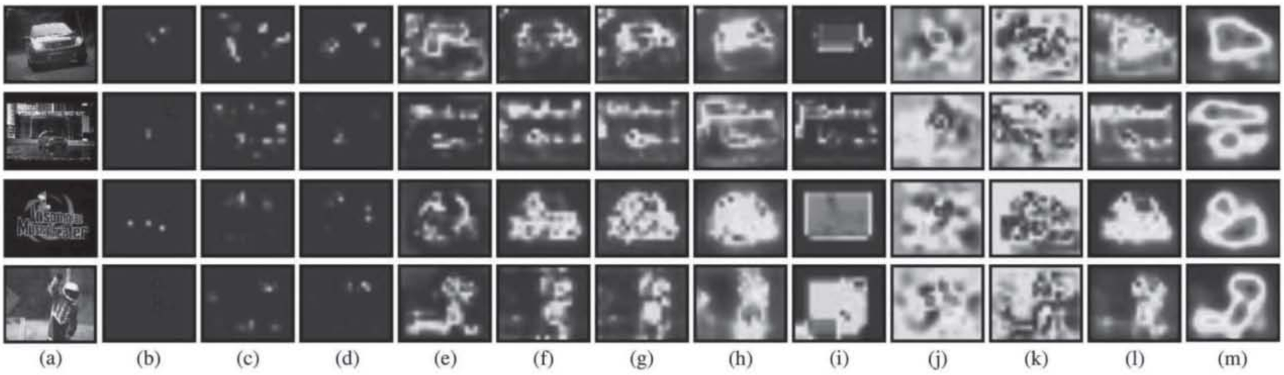
SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society ®

www.ieee.org/sp/index.html

APRIL 2010 VOLUME 17 NUMBER 4 ISPLEM
MAY 2010 VOLUME 17 NUMBER 5
JUNE 2010 VOLUME 17 NUMBER 6



Some representative results. Note that here the eye density maps are not convolved with a Gaussian kernel, which is a popular method to recover more positive samples for the evaluation. (a) Original frames; (b) eye fixation maps; (c) Itti98; (d) Itti01; (e) Itti05; (f) Hou07; (g) Guo08; (h) Harel07; (i) Zhai06; (j) Peters07; (k) Kienzle07; (l) Navalpakkam07; (m) our approach. For more see "Cost-Sensitive Rank Learning From Positive and Unlabeled Data for Visual Saliency Estimation," by Li *et al.*, p. 591.

For the April 2010 issue, see p. 325 for Table of Contents
For the May 2010 issue, see p. 413 for Table of Contents
For the June 2010 issue, see p. 525 for Table of Contents

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

JUNE 2010 VOLUME 17 NUMBER 6 ISPLEM (ISSN 1070-9908)

LETTERS

Biomedical Signal Processing

Digital and Multirate Signal Processing

Image and Multidimensional Signal Processing

Multimedia Signal Processing

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 $206.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2010 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage pending at New York, NY and at additional mailing offices. **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.Printed in U.S.A.

Sensor Array and Multichannel Signal Processing

Signal Processing for Communications

Speech Processing

Statistical and Adaptive Signal Processing

CORRESPONDENCE

Digital and Multirate Signal Processing

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society®

www.ieee.org/sp/index.html

MAY 2010 VOLUME 17 NUMBER 5 ISPLEM (ISSN 1070-9908)

LETTERS

Audio and Electroacoustics

Digital and Multirate Signal Processing

Image and Multidimensional Signal Processing

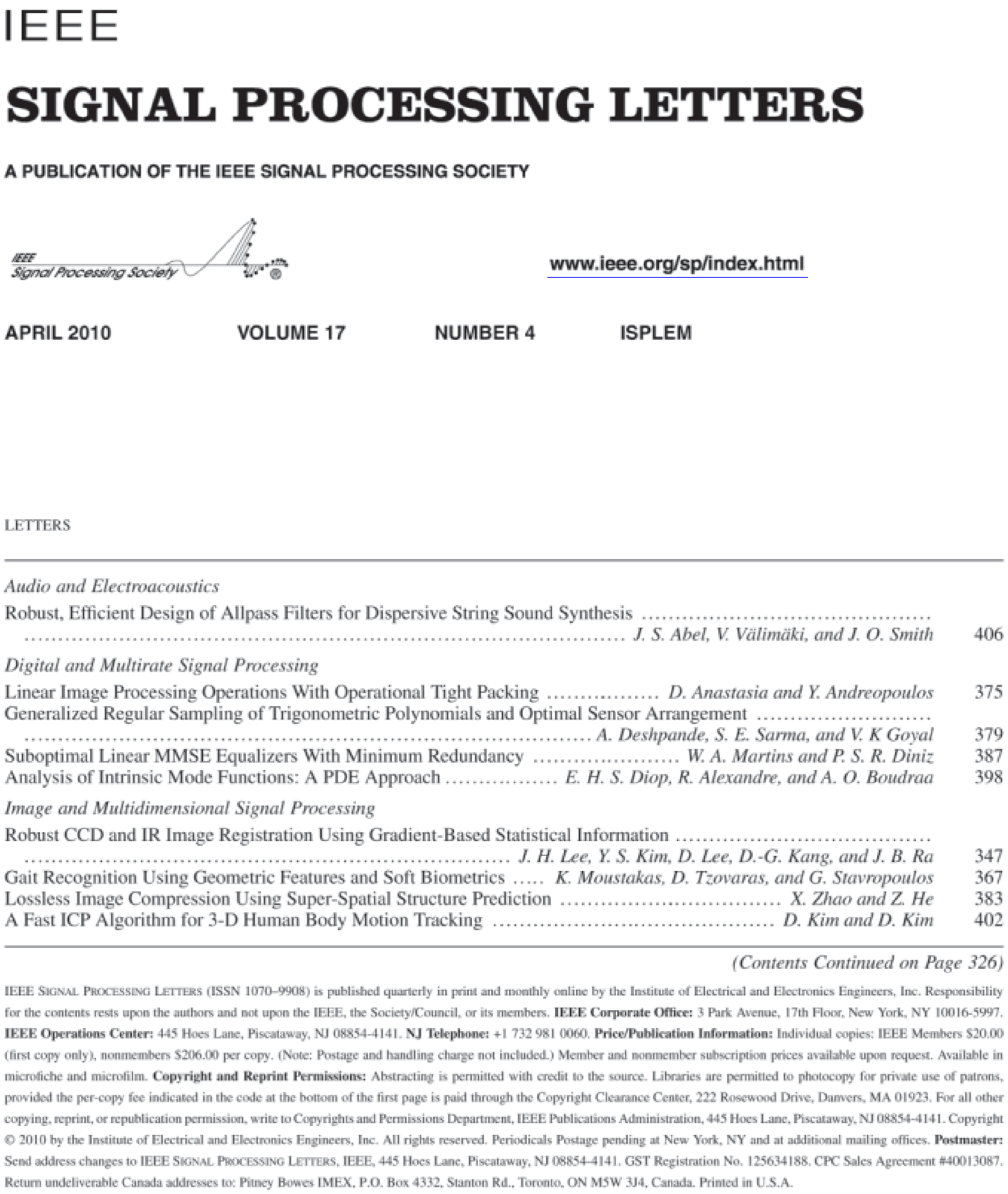
IEEE Signal Processing Letters (ISSN 1070–9908) is published quarterly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 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 $206.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2010 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage pending at New York, NY and at additional mailing offices. **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.Printed in U.S.A.

Sensor Array and Multichannel Signal Processing

Signal Processing for Communications

Speech Processing

Statistical and Adaptive Signal Processing



IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

APRIL 2010 VOLUME 17 NUMBER 4 ISPLEM

LETTERS

Audio and Electroacoustics

Digital and Multirate Signal Processing

Image and Multidimensional Signal Processing

(Contents Continued on Page 326)

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 $206.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Available in microfiche and microfilm. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2010 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals Postage pending at New York, NY and at additional mailing offices. **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. Printed in U.S.A.

(Contents Continued from Page 325)

Signal Processing for Communications

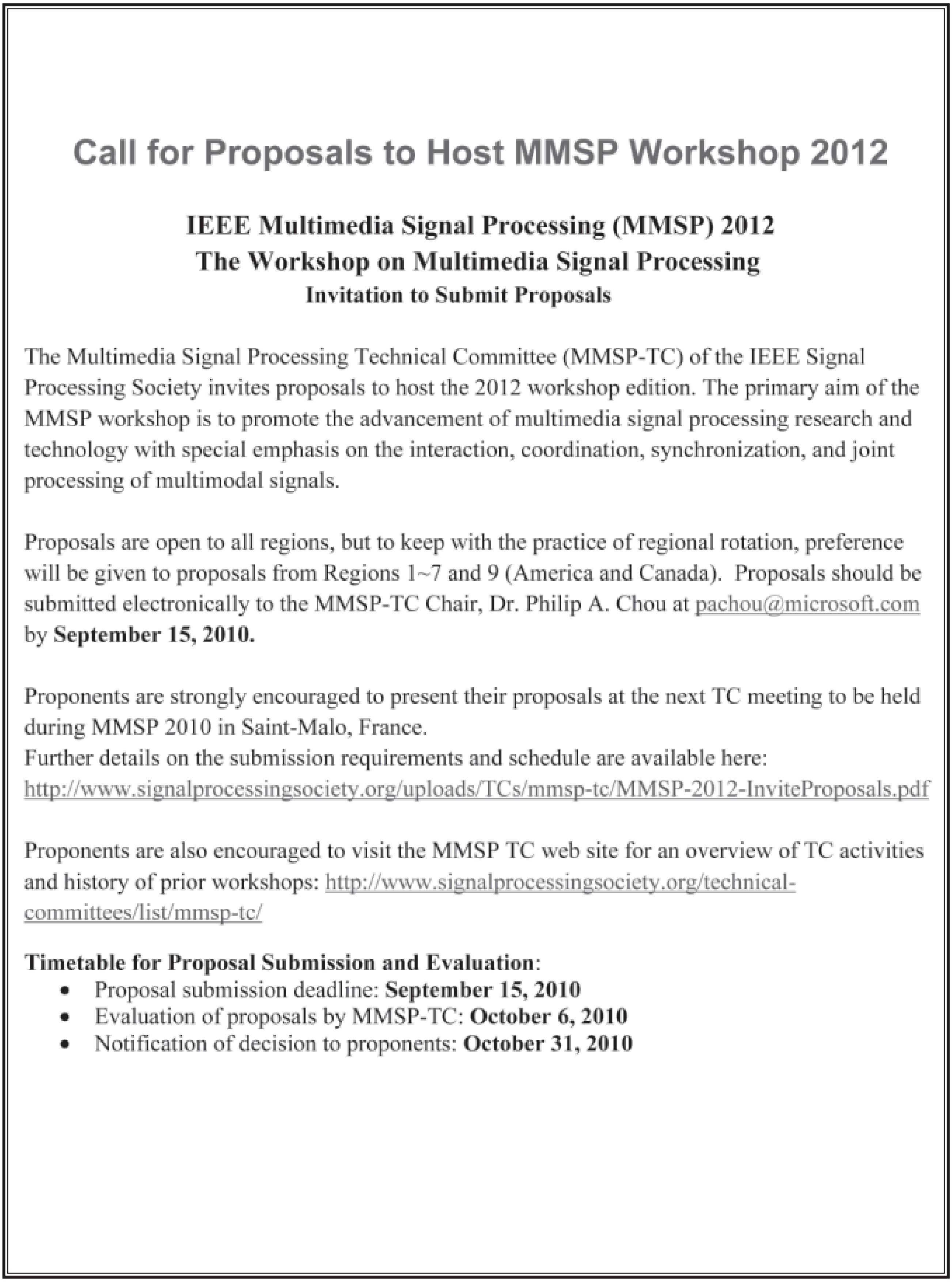
Statistical and Adaptive Signal Processing

CORRESPONDENCE

Statistical and Adaptive Signal Processing

Digital and Multirate Signal Processing

Signal Processing for Communications



Call for Proposals to Host MMSP Workshop 2012

IEEE Multimedia Signal Processing (MMSP) 2012
The Workshop on Multimedia Signal Processing
Invitation to Submit Proposals

The Multimedia Signal Processing Technical Committee (MMSP-TC) of the IEEE Signal Processing Society invites proposals to host the 2012 workshop edition. The primary aim of the MMSP workshop is to promote the advancement of multimedia signal processing research and technology with special emphasis on the interaction, coordination, synchronization, and joint processing of multimodal signals.

Proposals are open to all regions, but to keep with the practice of regional rotation, preference will be given to proposals from Regions 1~7 and 9 (America and Canada). Proposals should be submitted electronically to the MMSP-TC Chair, Dr. Philip A. Chou at pachou@microsoft.com by **September 15, 2010.**

Proponents are strongly encouraged to present their proposals at the next TC meeting to be held during MMSP 2010 in Saint-Malo, France.
Further details on the submission requirements and schedule are available here:
http://www.signalprocessingsociety.org/uploads/TCs/mmsp-tc/MMSP-2012-InviteProposals.pdf

Proponents are also encouraged to visit the MMSP TC web site for an overview of TC activities and history of prior workshops: http://www.signalprocessingsociety.org/technical-committees/list/mmsp-tc/

Timetable for Proposal Submission and Evaluation:

- Proposal submission deadline: **September 15, 2010**
- Evaluation of proposals by MMSP-TC: **October 6, 2010**
- Notification of decision to proponents: **October 31, 2010**

Call for Papers
IEEE Signal Processing Society
IEEE Journal of Selected Topics in Signal Processing

Special Issue on Signal Processing in Gossiping Algorithms Design and Applications

Distributed consensus and gossiping algorithms have recently spurred a number of new research results on decentralized signal processing. In this context, network gossiping models can solve a variety of classical sensor array processing, fusing data collected at sensors that are not co-located to perform adaptive filtering, computing parameter estimates or decisions, without central coordination. Closely related to these studies is the research focused on optimizing objective functions in a decentralized setting. Given the importance gained by the network gossiping primitive, many new studies have analyzed the tradeoffs that exist between computation accuracy and communication cost, particularly in wireless media, and the impact of fading and mobility on the information diffusion speed. Gossiping has also been used as a tool to compress data as they are aggregated. This special issue intends to attract papers that advance fundamentally the application and understanding of the network gossiping primitive for decentralized signal processing over wireless sensor networks.

We invite original and unpublished research contributions in all areas relevant to decentralized signal processing through network gossiping. The topics of interest include, but are not limited to:

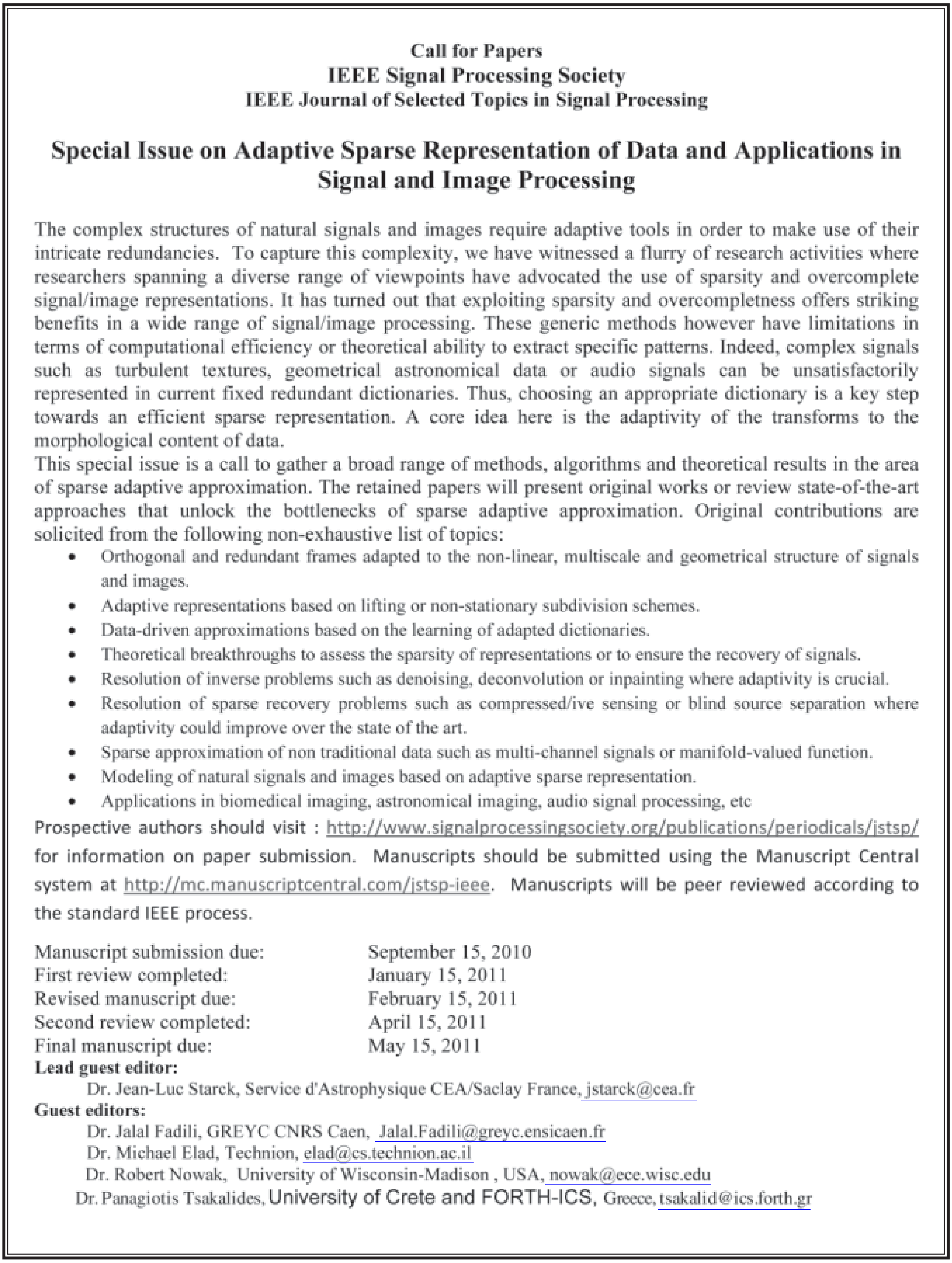
- Decentralized estimation and detection algorithms via network gossiping
- Gossip algorithms for sensor fusion and querying
- Compressive data aggregation via gossip algorithms
- Performance analysis of gossiping protocols and scaling laws
- Decentralized optimization via network gossiping
- Gossip algorithms for decentralized adaptive filtering
- Gossiping models for social networks analysis and social learning
- Gossip algorithms performance under communication constraints
- Coding and channel access methods for wireless network gossiping

Prospective authors should visit http://www.signalprocessingsociety.org/publications/periodicals/jstsp/ for information on paper submission. Manuscripts should be submitted using the Manuscript Central system at http://mc.manuscriptcentral.com/jstsp-ieee. Manuscripts will be peer reviewed according to the standard IEEE process.

Manuscript submission due:	**June 20, 2010**
First review completed:	Sept 20, 2010
Revised manuscript due:	Oct 30, 2010
Second review completed:	Dec 30, 2010
Final manuscript due:	Jan 30, 2011

Lead guest editor:
Anna Scaglione, UC Davis, USA, ascaglione@ucdavis.edu

Guest editors:
Mark Coates, McGill University, Canada, coates@ece.mcgill.ca
Michael Gastpar, UC Berkeley, USA, gastpar@eecs.berkeley.edu
John Tsitsiklis, MIT, USA jnt@mit.edu
Martin Vetterli, EPFL, Switzerland, martin.vetterli@epfl.ch



Call for Papers
IEEE Signal Processing Society
IEEE Journal of Selected Topics in Signal Processing

Special Issue on Adaptive Sparse Representation of Data and Applications in Signal and Image Processing

The complex structures of natural signals and images require adaptive tools in order to make use of their intricate redundancies. To capture this complexity, we have witnessed a flurry of research activities where researchers spanning a diverse range of viewpoints have advocated the use of sparsity and overcomplete signal/image representations. It has turned out that exploiting sparsity and overcompletness offers striking benefits in a wide range of signal/image processing. These generic methods however have limitations in terms of computational efficiency or theoretical ability to extract specific patterns. Indeed, complex signals such as turbulent textures, geometrical astronomical data or audio signals can be unsatisfactorily represented in current fixed redundant dictionaries. Thus, choosing an appropriate dictionary is a key step towards an efficient sparse representation. A core idea here is the adaptivity of the transforms to the morphological content of data.

This special issue is a call to gather a broad range of methods, algorithms and theoretical results in the area of sparse adaptive approximation. The retained papers will present original works or review state-of-the-art approaches that unlock the bottlenecks of sparse adaptive approximation. Original contributions are solicited from the following non-exhaustive list of topics:

- Orthogonal and redundant frames adapted to the non-linear, multiscale and geometrical structure of signals and images.
- Adaptive representations based on lifting or non-stationary subdivision schemes.
- Data-driven approximations based on the learning of adapted dictionaries.
- Theoretical breakthroughs to assess the sparsity of representations or to ensure the recovery of signals.
- Resolution of inverse problems such as denoising, deconvolution or inpainting where adaptivity is crucial.
- Resolution of sparse recovery problems such as compressed/ive sensing or blind source separation where adaptivity could improve over the state of the art.
- Sparse approximation of non traditional data such as multi-channel signals or manifold-valued function.
- Modeling of natural signals and images based on adaptive sparse representation.
- Applications in biomedical imaging, astronomical imaging, audio signal processing, etc

Prospective authors should visit : http://www.signalprocessingsociety.org/publications/periodicals/jstsp/ for information on paper submission. Manuscripts should be submitted using the Manuscript Central system at http://mc.manuscriptcentral.com/jstsp-ieee. Manuscripts will be peer reviewed according to the standard IEEE process.

Manuscript submission due:	September 15, 2010
First review completed:	January 15, 2011
Revised manuscript due:	February 15, 2011
Second review completed:	April 15, 2011
Final manuscript due:	May 15, 2011

Lead guest editor:
Dr. Jean-Luc Starck, Service d'Astrophysique CEA/Saclay France, jstarck@cea.fr

Guest editors:
Dr. Jalal Fadili, GREYC CNRS Caen, Jalal.Fadili@greyc.ensicaen.fr
Dr. Michael Elad, Technion, elad@cs.technion.ac.il
Dr. Robert Nowak, University of Wisconsin-Madison , USA, nowak@ece.wisc.edu
Dr. Panagiotis Tsakalides, University of Crete and FORTH-ICS, Greece, tsakalid@ics.forth.gr

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society®

www.ieee.org/sp/index.html

JUNE 2010 VOLUME 4 NUMBER 3 IJSTGY (ISSN 1932-4553)

General Model Order Selection Theories

Model Order Determination for Applied Statistics

IEEE

Model Order Determination for Practical Applications

Miscellaneous

ANNOUNCEMENTS

Call for Papers
IEEE Journal of Selected Topics in Signal Processing
Special Issue on Music Signal Processing

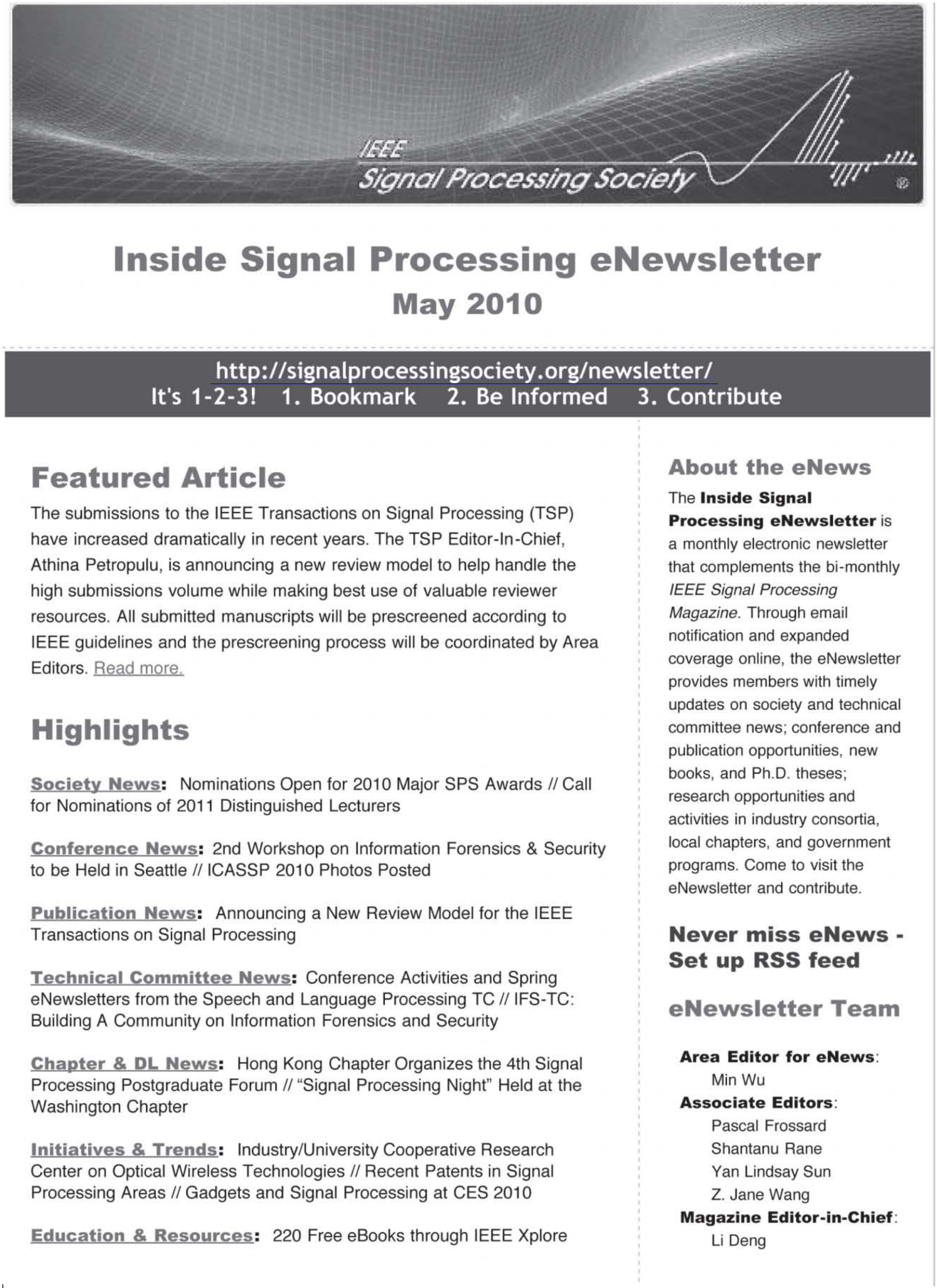
The extraction of meaningful information from audio waveform data is a central application of digital signal processing. When dealing with specific audio domains such as speech or music, it is crucial to properly understand and apply the appropriate domain-specific properties, be they acoustic, linguistic, or musical. This special issue seeks to gather contributions that address aspects of music signal processing by explicitly incorporating the distinctive characteristics of music audio, such as the presence of multiple, coordinated sources, the existence of structure at many temporal levels, and the peculiar kinds of information being carried.

Of particular interest are papers that aim to establish a rigorous foundation for the processing of music signals, in contrast to the widespread approach of borrowing and adapting techniques from speech processing. For example, new signal models and parameter estimation algorithms are required that can accommodate polyphonic and multitimbral music signals. Furthermore, in the design of musically-meaningful audio features, approaches must account for high-level musical aspects regarding melody, harmony, rhythm, and other acoustic and structural properties. All contributions should have a clear focus on the processing of waveform-based music audio recordings. However, the work may also exploit complementary sources of musical information such as lyrics, symbolic score data, MIDI, or textual annotations.

The complexity and diversity of music data makes automatic music signal processing a challenging field of research. Music processing systems may need to take account of aspects such as genre (e. g., pop music, classical music, jazz, ethnological music), instrumentation (e. g., orchestra, piano, drums, voice), and many other musical properties (e. g., dynamics, tempo, or timbre). Our goal for this special issue is to spur progress in core techniques needed for the future signal processing systems that will enable users to access and explore music in all its different facets.

Submission deadline: **September 20, 2010**
First Review completed: December 20, 2010
Revised manuscript due: February 20, 2011
Second Review completed: March 20, 2011
Final manuscript due: April 10, 2011
Tentative publication date: May 2011

Guest Editors:
Dr. Meinard Muller, Saarland University and MPI Informatik, Germany, meinard@mpi-inf.mpg.de
Dr. Shigeki Sagayama, The University of Tokyo, Japan, sagayama@hil.t.u-tokyo.ac.jp
Dr. Anssi Klapuri, Queen Mary University of London, UK, anssi.klapuri@elec.qmul.ac.uk
Dr. Gael Richard, T´el´ecom ParisTech, France, gael.richard@enst.fr
Dr. Daniel Ellis, Columbia University, USA, dpwe@ee.columbia.edu



Inside Signal Processing eNewsletter

May 2010

http://signalprocessingsociety.org/newsletter/
It's 1-2-3! 1. Bookmark 2. Be Informed 3. Contribute

Featured Article

The submissions to the IEEE Transactions on Signal Processing (TSP) have increased dramatically in recent years. The TSP Editor-In-Chief, Athina Petropulu, is announcing a new review model to help handle the high submissions volume while making best use of valuable reviewer resources. All submitted manuscripts will be prescreened according to IEEE guidelines and the prescreening process will be coordinated by Area Editors. Read more.

Highlights

Society News: Nominations Open for 2010 Major SPS Awards // Call for Nominations of 2011 Distinguished Lecturers

Conference News: 2nd Workshop on Information Forensics & Security to be Held in Seattle // ICASSP 2010 Photos Posted

Publication News: Announcing a New Review Model for the IEEE Transactions on Signal Processing

Technical Committee News: Conference Activities and Spring eNewsletters from the Speech and Language Processing TC // IFS-TC: Building A Community on Information Forensics and Security

Chapter & DL News: Hong Kong Chapter Organizes the 4th Signal Processing Postgraduate Forum // "Signal Processing Night" Held at the Washington Chapter

Initiatives & Trends: Industry/University Cooperative Research Center on Optical Wireless Technologies // Recent Patents in Signal Processing Areas // Gadgets and Signal Processing at CES 2010

Education & Resources: 220 Free eBooks through IEEE Xplore

About the eNews

The **Inside Signal Processing eNewsletter** is a monthly electronic newsletter that complements the bi-monthly *IEEE Signal Processing Magazine*. Through email notification and expanded coverage online, the eNewsletter provides members with timely updates on society and technical committee news; conference and publication opportunities, new books, and Ph.D. theses; research opportunities and activities in industry consortia, local chapters, and government programs. Come to visit the eNewsletter and contribute.

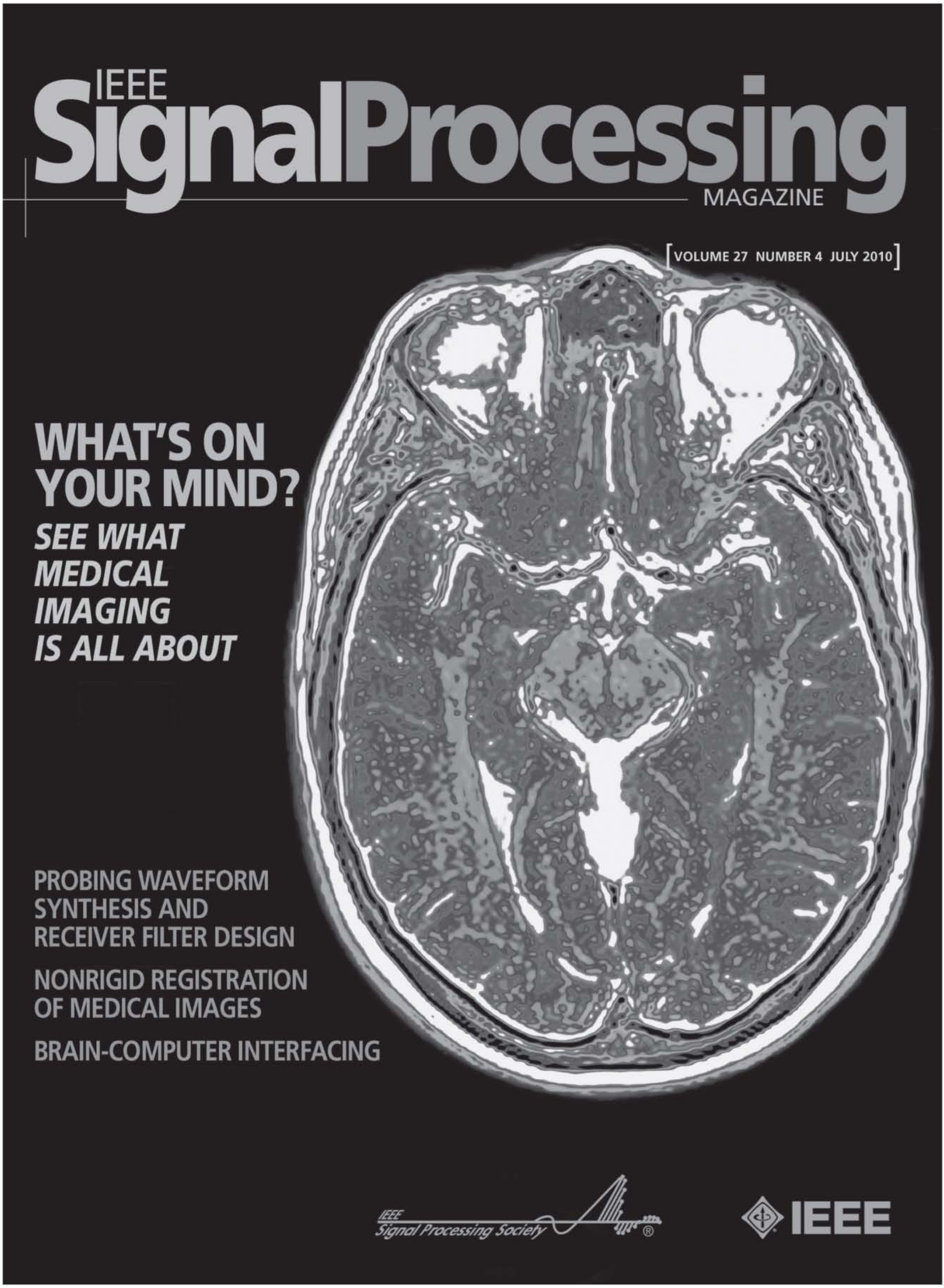
Never miss eNews - Set up RSS feed

eNewsletter Team

Area Editor for eNews:
Min Wu

Associate Editors:
Pascal Frossard
Shantanu Rane
Yan Lindsay Sun
Z. Jane Wang

Magazine Editor-in-Chief:
Li Deng



IEEE SignalProcessing MAGAZINE

[VOLUME 27 NUMBER 4 JULY 2010]

WHAT'S ON YOUR MIND?

SEE WHAT MEDICAL IMAGING IS ALL ABOUT

PROBING WAVEFORM SYNTHESIS AND RECEIVER FILTER DESIGN

NONRIGID REGISTRATION OF MEDICAL IMAGES

BRAIN-COMPUTER INTERFACING

IEEE Signal Processing Society

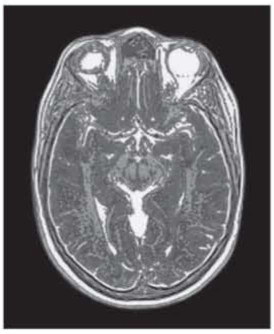
IEEE

[VOLUME 27 NUMBER 4]

[CONTENTS]

[SPECIAL SECTION—SIGNAL AND IMAGE PROCESSING IN MEDICAL IMAGING]

[COVER] © PHOTODISC/DON FARRALL



[FEATURE]

[COLUMNS]

[DEPARTMENT]

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.

IEEE SIGNAL PROCESSING MAGAZINE 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 $20.00 (first copy only), nonmembers $175.00 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. Copyright©2010 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. 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.2010.936720

CALL FOR PAPERS
IEEE Transactions on Multimedia
Special Issue on Interactive Multimedia

Schedule:	Guest Editors:
Manuscript submission: 1 September 2010 Acceptance/Revision notification: 1 January 2011 Revised manuscript due: 15 February 2011 Final acceptance notification: 1 April 2011 Final manuscript due: 15 April 2011 Tentative publication: August 2011	Prof. S.-H. Gary Chan, HKUST, Hong Kong (gchan@cse.ust.hk) Dr. Jin Li, Microsoft Research, Microsoft Research, U.S.A. (jinl@microsoft.com) Prof. Pascal Frossard, EPFL, Switzerland (pascal.frossard@epfl.ch) Dr. Gerasimos Potamianos, NCSR "Demokritos" (gpotam@iit.demokritos.gr)

With the advances in broadband networks, networking and QoS standards, audio/video coding and processing techniques and multimedia-capable user devices, multimedia streaming over networks has become a reality. With the popularity of peer-to-peer and social network applications, there has been increasing interest of interactive multimedia applications. An interactive multimedia application refers to live sharing of multimedia contents in terms of video, audio, texts or images among distributed users in a network. An interactive session requires real-time processing of data and media streams, with the support of user interactions at any time. Examples are voice over IP (VoIP), video conferencing, distributed collaborative environments, teleconferencing, online multiplayer games, social games, etc. As enterprises and interpersonal/business communications are increasingly global, such distributed interactive multimedia applications overcome accessibility and co-location barriers by bringing people together, leading to tremendous saving in time, operational and fuel costs.

Distributed interactive multimedia application is one of the fastest growing market sectors. While there are many business opportunities and advancements, the design of a good interactive multimedia system still faces many technological challenges today. Overcoming these challenges requires joint effort of various multimedia communities of system integration and architecture, signal processing, communication/coding and transmission, network design and measurement, standardization, etc. Furthermore, design of good user interfaces for smart interactive systems, and the incorporation of automatic perception of human activity (presence, speech, interaction), remains an important area at its infancy.

This special issue intends to bring together papers from experts in various multimedia areas to address challenges and present effective solutions for interactive multimedia applications, as well as to promote the development of novel interactive technologies. We solicit original contributions in the areas related to, but not limited to, the following:

- Multimedia processing for interactive applications
  - Scheduling and coding techniques for interactive VoIP and multimedia conferencing
  - Congestion control and QoS/error correction to mitigate network anomalies
  - Interactive multimedia messaging protocol
  - Low bit-rate multimedia processing and delivery
- Design of collaborative conferencing networks
  - Novel architecture and optimization for interactive multimedia applications
  - Real-time, low-delay and interactive telepresence networks
  - Interactive technologies and applications over mobile, ad-hoc or infrastructure-based overlay (peer-to-peer) or social networks
- Support of real-time interactivity
  - Session initiation, maintenance and control
  - Quality monitoring and management for multi-party voice and video interactive applications
  - Security and privacy solutions
- Multimodal perception technologies of human activity in the design of smart interactive systems
  - Design of smart spaces for interactive systems
  - Automatic detection of human presence and speech
  - Speech enhancement and transcription, speaker localization and interactive control of audio-visual content
  - User interface design
- Measurements and standards for interactive applications
  - System design, integration, trials and measurements
  - Success or failure experiences for interactive multimedia systems or networks
  - Standardization activities for interactive multimedia

Papers should be formatted according to the IEEE Transactions on Multimedia guidelines for authors (see: http://www.ieee.org/organizations/society/tmm/author_info.html). Mandatory overlength page charges and color charges will apply. Manuscripts (both 1-column and 2-column versions are required) should be submitted electronically through the online IEEE manuscript submission system at http://tmmieee.manuscriptcentral.com/. When selecting a manuscript type, authors must click on Special Issue on Interactive Multimedia. A copyright form with the manuscript number on the top of the page is required to be completed, signed and faxed to 1-732-562-8905 at the time of submission.

IEEE International Conference on Acoustics, Speech and Signal Processing

ICASSP 2016 Call for Proposal

Are you interested in hosting an ICASSP conference?

Proposals will be presented at ICASSP 2011 in Prague. Czech Republic.

Future sites for ICASSP should be selected four to five years in advance.

A Signal Processing Society member who is interested in hosting an ICASSP must submit a proposal to the Conference Board six (6) months prior to the next Conference Board meeting at ICASSP.

For complete information please visit:
http://www.signalprocessingsociety.org/uploads/conferences/ConferenceOrganizersHandbook.pdf

IEEE Signal Processing Society Transactions
Information for Authors

The IEEE TRANSACTIONS are published covering advances in the theory and application of signal processing. The scopes are reflected in the EDICS: the Editor's Information and Classification Scheme. Please consider the journal with the most appropriate scope for your submission.
Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Correspondences (brief items that describe a use for or magnify the meaning of a single technical point, or provide comment on a paper previously published in the TRANSACTIONS). Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.
Every manuscript must (a) provide a clearly defined statement of the problem being addressed, (b) state why it is important to solve the problem, and (c) give an indication as to how the current solution fits into the history of the problem.
By submission/resubmission of your manuscript to these TRANSACTIONS, you are acknowledging that you 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.
New and revised manuscripts should be prepared following the "New Manuscript Submission" guidelines below, and submitted to the online manuscript system Manuscript Central. After acceptance, finalized manuscripts should be prepared following the "Final Manuscript Submission Guidelines" below. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will access your manuscript electronically via the Manuscript Central system.

New Manuscript Submission. Please follow the next steps.
1. *Account in Manuscript Central.* If necessary, create an account in the on-line manuscript system Manuscript Central. 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-spaced format (one full blank line between lines of type) using a font size of 11 points or larger, having a margin of at least 1 inch on all sides. For a regular paper, the manuscript may not exceed 30 double-spaced pages, including title; names of authors and their complete contact information; abstract; text; all images, figures and tables; and all references. Overlength page charges are levied beginning with the 11th published page of the manuscript. You are, therefore, advised to be conservative in your submission.
Upload your manuscript as a PDF file "manuscript.pdf" to the Manuscript Central web site; then proofread your submission, confirming that all figures and equations are visible in your document before pressing the button. Proofreading is critical; once you press the button, your manuscript cannot be changed in any way. You may also submit your manuscript as a Post-Script 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. *Additional Material for Review.* Please upload pdf versions of all items in the reference list which are not publicly available, such as unpublished (submitted) papers. Other materials for review such as supplementary tables and figures, audio fragments and quicktime movies may be uploaded as well. Reviewers will be able to view these files only if they have the appropriate software on their computers. 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.
4. *Double-Column Version of Manuscript.* You are required to also submit a roughly formatted version of the manuscript in single-spaced, double column IEEE format (10 points for a regular submission or 9 points for a Correspondence) using the IEEE style files (it is allowed to let long equations stick out). This version will serve as a confirmation of the approximate publication length of the manuscript at submission, and gives an additional confirmation of your understanding that overlength page charges will be paid hen billed. Upload this version of the manuscript as a PDF file "double.pdf" to the Manuscript Central web site.
5. *Submission.* After uploading all files and proofreading them, submit your manuscript by pressing the button. 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 press the button, your manuscript cannot be changed in any way.
6. *Copyright 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 authors. *How to submit the properly executed and signed the electronic copyright Form (eCF).* IEEE policy requires that prior to publication all authors or their employers must transfer to the IEEE in writing any copyright they hold for their individual papers. When you are redirected to the IEEE Electronic Copyright Form wizard at the end of your original submission, please sign the eCF by typing your name at the proper location and click on a "Submit" button. You may also print the form available at http://www.ieee.org/web/publications/rights/index.html and fax it to the IEEE Signal Processing Society Publications Office at +1 732 562 8905.

Correspondence Items. Correspondence items are short disclosures with a reduced scope or significance that typically describe a use for or magnify the meaning of a single technical point, or provide brief comments on material previously published in the TRANSACTIONS. These items may not exceed 12 pages in double-spaced format (3 pages for Comments), using 11 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 "New Manuscript Submission" above for instructions).

Manuscript Length. Papers published on or after 1 January 2007 can now be up to 10 pages, and any paper in excess of 10 pages will be subject to overlength page charges. The IEEE Signal Processing Society has determined that the standard manuscript length shall be no more than 10 published pages (double-column format, 10 point type) for a regular submission, or 6 published pages (9 point type) for a Correspondence item, respectively. Manuscripts that exceed these limits will incur mandatory overlength page charges, as discussed below. 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. Exceptions to the 30-page (regular paper) or 12-page (Correspondences) manuscript length 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 rejected manuscripts are allowed to resubmit their manuscripts only once. The Society strongly discourages resubmission of rejected manuscripts more than once. At the time of submission, you will be asked whether you consider your manuscript to be a new submission or a resubmission of an earlier rejected manuscript.

If you choose to submit afresh a new version of your manuscript, you will be asked to submit a supporting document detailing how your new version addresses all of the reviewers' comments. You may request at the time of submission that the same Associate Editor handle your manuscript. The Editor in Chief will do his/her best effort to accommodate your request while taking into consideration the balancing of the workload among the AEs of the editorial board of the journal. The new manuscript, the old manuscript, the reviews, and the supporting document detailing your response will be made available by the Associate Editor to the reviewers of your new version.

Full details of the resubmission process can be found in the Signal Processing Society "Policy and Procedures Manual" at http://www.signalprocessingsociety.org/about/governance/policy-procedure/

Author Misconduct.
Author Misconduct Policy: 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 authors previously published work that also involves other authors. Plagiarism is unacceptable. Self-plagiarism involves the verbatim copying or reuse of an author's own prior work without appropriate citation; it is also unacceptable. Self-plagiarism includes 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. Authors may only submit original work that has not appeared elsewhere in a journal publication, nor is under review for another journal publication. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper. If authors have used their own previously published work as a basis for a new submission, they are required to cite the previous work and very briefly indicate how the new submission offers substantively novel contributions beyond those of the previously published 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 related prior work; the papers cannot be identical; and the journal publication must include novel aspects.
Author Misconduct Procedures: The procedures that will be used by the Signal Processing Society in the investigation of author misconduct allegations are described in the IEEE SPS Policies and Procedures Manual.
Author Misconduct Sanctions: The IEEE Signal Processing Society will apply the following sanctions in any case of plagiarism, or in cases of self-plagiarism that involve an overlap of more than 25% with another journal manuscript:
1) Immediate rejection of the manuscript in question;

2) Immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors;
3) Prohibition against each of the authors for any new submissions, either individually, in combination with the authors of the plagiarizing manuscript, or in combination with new co-authors, to all of the Society's publications (journals, conferences, workshops). The prohibition shall continue for one year from notice of suspension. Further, plagiarism and self-plagiarism may also be actionable by the IEEE under the rules of Member Conduct.

Submission Format.
Authors are encouraged to prepare manuscripts employing the on-line style files developed by IEEE. 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
http://www.ieee.org/web/publications/authors/transjnl/index.html
(LaTeX and MS Word).
Authors using LaTeX: the two PDF versions of the manuscript needed for submission can both be produced by the IEEEtran.cls style file.
A double-spaced document is generated by including \documentclass[11pt,draftcls,onecolumn] {IEEEtran} as the first line of the manuscript source file, and a single-spaced double-column document for estimating the publication page charges via \documentclass[10pt,twocolumn,twoside]{IEEEtran} for a regular submission, or \documentclass[9pt,twocolumn,twoside]{IEEEtran} for a Correspondence item.
• *Title page and abstract:* The first page of the manuscript shall contain the title, names and contact information for all authors (full mailing address, institutional affiliations, phone, fax, and e-mail), the abstract, and the EDICS. An asterisk * should be placed next to the name of the Corresponding Author who will serve as the main point of contact for the manuscript during the review and publication processes.
An abstract should have not more than 200 words for a regular paper, or 50 words for a Correspondence item. The abstract should indicate the scope of the paper or Correspondence, and summarize the author's conclusions. This will make the abstract, by itself, a useful tool for information retrieval.
• *EDICS:* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at http://ewh.ieee.org/soc/sps/tsp/.
The EDICS' category should appear on the first page—i.e., the title and abstract page—of the manuscript.
• *Illustrations and tables:* Each figure and table should have a caption that is intelligible without requiring reference to the text. Illustrations/tables may be worked into the text of a newly-submitted manuscript, or placed at the end of the manuscript. (However, for the final submission, illustrations/tables must be submitted separately and not interwoven with the text.)
Illustrations in color may be used but, unless the final publishing will be in color, the author is responsible that the corresponding grayscale figure is understandable. In preparing your illustrations, note that in the printing process, most illustrations are reduced to single-column width to conserve space. This may result in as much as a 4:1 reduction from the original. Therefore, make sure that all words are in a type size that will reduce to a minimum of 9 points or 3/16 inch high in the printed version. Only the major grid lines on graphs should be indicated.
• *Abbreviations:* This TRANSACTIONS follows the practices of the IEEE on units and abbreviations, as outlined in the Institute's published standards. See http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/authors/ transjnl/auinfo07.pdf for details.
• *Mathematics:* All mathematical expressions must be legible. Do not give derivations that are easily found in the literature; merely cite the reference.

Final Manuscript Submission Guidelines.
Upon formal acceptance of a manuscript for publication, instructions for providing the final materials required for publication will be sent to the Corresponding Author. Finalized manuscripts should be prepared in LaTeX or MS Word, and are required to use the style files established by IEEE, available at
http://www.ieee.org/web/publications/authors/transjnl/index.html.
Instructions for preparing files for electronic submission are as follows:
• Files must be self-contained; that is, there can be no pointers to your system setup.
• Include a header to identify the name of the TRANSACTIONS, the name of the author, and the software used to format the manuscript.
• Do not import graphics files into the text file of your finalized manuscript (although this is acceptable for your initial submission). If submitting on disk, use a separate disk for graphics files.
• Do not create special macros.
• Do not send PostScript files of the text.
• File names should be lower case.
• Graphics files should be separate from the text, and not contain the caption text, but include callouts like "(a)," "(b)."
• Graphics file names should be lower case and named fig1.eps, fig2.tif, etc.
• Supported graphics types are EPS, PS, TIFF, or graphics created using Word, Powerpoint, Excel or PDF. Not acceptable is GIF, JPEG, WMF, PNG, BMP or any other format (JPEG is accepted for author photographs only). The provided resolution needs to be at least 600 dpi (400 dpi for color).
• Please indicate explicitly if certain illustrations should be printed in color; note that this will be at the expense of the author. Without other indications, color graphics will appear in color in the online version, but will be converted to grayscale in the print version. Additional instructions for preparing, verifying the quality, and submitting graphics are available via http://www.ieee.org/web/publications/authors/transjnl/index.html.

Multimedia Materials.
IEEE Xplore can publish multimedia files and 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 http://www.ieee.org/web/publications/authors/transjnl/index.html under "Multimedia." To make your work reproducible by others, the Transactions encourages you to submit all files that can recreate the figures in your paper.

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 (six 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. 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.
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 in print is expensive, and all charges for color are the responsibility of the author. The estimated costs are as follows. There will be a charge of $62.50 for each figure; this charge may be subject to change without notification. In addition, there are printing preparation charges which may be estimated as follows: color reproductions on four or fewer pages of the manuscript: a total of approximately $1045; color reproductions on five pages through eight pages: a total of approximately $2090; color reproductions on nine through 12 pages: a total of approximately $3135, and so on. Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in the TRANSACTIONS is considered acceptance of this requirement.

To find the Information for Authors for IEEE Signal Processing Letters or the IEEE Signal Processing Magazine, please refer to the IEEE Signal Processing website at www.signalprocessingsociety.org.

IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print in easy, cost-effective and quick.

Complete this form, tear it out, and either fax it (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
Box 1331
Piscataway, NJ 08855-1331
Call Reprint Department at (732) 562-3941/3917
for questions regarding this form
(732) 981-8062 - FAX

PLEASE SEND ME

□ 50 □ 100 □ 200 □ 300 □ 400 □ 500 or ________ (in multiples of 50) reprints.

□ YES □ NO Self-covering/title page required. COVER PRICE: $72 per 100, $37 per 50.

□ $56 Air Freight must be added for all orders being shipped outside the U.S.

□ $18.50 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: ______________

□ I have special instructions noted on the reverse side.

2009 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	$120	$201	$231	$238	$285	$340	$371	$408	$440	$477	$510	$543
100	$231	$403	$454	$478	$570	$680	$742	$817	$885	$953	$1021	$1088

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:______________

2010 IEEE SIGNAL PROCESSING SOCIETY MEMBERSHIP APPLICATION

(Current and reinstating IEEE members joining SPS complete areas 1, 2, 8, 9.)
Mail to: **IEEE OPERATIONS CENTER, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA**
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: SEND MAIL TO: □ **Home Address** OR □ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

TITLE FIRST OR GIVEN NAME MIDDLE NAME SURNAME/LAST NAME

HOME ADDRESS

CITY STATE/PROVINCE POSTAL CODE COUNTRY

2.

Are you now or were you ever a member of IEEE? □ **Yes** □ **No**
If yes, please provide, if known:

MEMBERSHIP NUMBER | | | | | | | | |

Grade **Year Membership Expired:**

3. BUSINESS/PROFESSIONAL INFORMATION

Company Name

Department/Division

Title/Position **Years in Current Position**

Years in the Profession Since Graduation □ **PE State/Province**

Street Address

City **State/Province** **Postal Code** **Country**

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A.

Baccalaureate Degree Received **Program/Course of Study**

College/University **Campus**

State/Province **Country** **Mo./Yr. Degree Received**

B.

Highest Technical Degree Received Program/Course of Study

College/University Campus

State/Province Country Mo./Yr. Degree Received

5.

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth
Day Month Year

□ Male □ Female

7. CONTACT INFORMATION

Office Phone/Office Fax Home Phone/Home Fax

Office E-Mail Home E-Mail

8. 2010 IEEE MEMBER RATES

IEEE DUES Residence	16 Aug 08-28 Feb 09 Pay Full Year Renewing	 Joining	1 Mar 09-15 Aug 09 Pay Half Year** Renewing	 Joining
Region 1	$180.00 □	$175.00□	$ 90.00 □	$ 88.00 □
Region 2	$177.00 □	$175.00□	$ 89.00 □	$ 88.00 □
Region 3	$177.00 □	$175.00□	$ 89.00 □	$ 88.00 □
Region 3 (Jamaica)	$136.00 □	$134.00□	$ 68.00 □	$ 67.00 □
Region 4	$178.00 □	$175.00□	$ 89.00 □	$ 88.00 □
Region 5	$177.00 □	$175.00□	$ 90.00 □	$ 88.00 □
Region 6	$176.00 □	$175.00□	$ 88.00 □	$ 88.00 □
Region 7 (GST)	$159.70 □	$159.70□	$ 79.90 □	$ 79.90 □
Region 7 (HST)	$170.42 □	$170.42□	$ 85.20 □	$ 85.20 □
Region 8	$147.00 □	$147.00□	$ 74.00 □	$ 74.00 □
Region 9	$138.00 □	$138.00□	$ 69.00 □	$ 69.00 □
Region 10	$139.00 □	$139.00□	$ 70.00 □	$ 70.00 □
Region 10 (Japan Sections)	$164.00 □	$139.00□	$ 82.00 □	$ 70.00 □

9. 2010 SPS MEMBER RATES

		16 Aug.-28 Feb. Pay Full Year	1 Mar.-15 Aug. Pay Half Year**
Signal Processing Society Membership Fee*		$ 29.00 □	$ 14.50 □

SPS Digital Library (included in membership: electronic access to IEEE Signal Processing Magazine, IEEE Signal Processing Letters, IEEE Journal of Selected Topics in Signal Processing, IEEE Trans. on Audio, Speech, and Language Processing, IEEE Trans. on Image Processing, IEEE Trans. on Information Forensics and Security and IEEE Trans. on Signal Processing)

Publications available only with SPS membership:

		16 Aug.-28 Feb. Pay Full Year	1 Mar.-15 Aug. Pay Half Year**
Image Processing, IEEE Transactions on:	Print	$ 77.00 □	$ 39.00 □
Signal Processing, IEEE Transactions on:	Print	$ 69.00 □	$ 35.00 □
Audio, Speech, and Language Processing, IEEE Transactions on:	Print	$ 45.00 □	$ 23.00 □
Information Forensics and Security, IEEE Transactions on:	Print	$ 48.00 □	$ 24.00 □
IEEE Journal of Selected Topics in Signal Processing:	Print	$ 69.00 □	$ 35.00 □
Computing in Science & Engineering Magazine:	Combo-Print & Electronic	$ 47.00 □	$ 24.00 □
Medical Imaging, IEEE Transactions on:	Print	$ 70.00 □	$ 35.00 □
	Electronic	$ 50.00 □	$ 25.00 □
	Combo-Print & Electronic	$ 85.00 □	$ 43.00 □
Mobile Computing, IEEE Transactions on:	Combo-Print & Electronic	$ 48.00 □	$ 24.00 □
Multimedia, IEEE Transactions on:	Combo-Print & Electronic	$ 50.00 □	$ 25.00 □
IEEE MultiMedia Magazine:	Combo-Print & Electronic	$ 40.00 □	$ 20.00 □
IEEE Security and Privacy Magazine:	Combo-Print and Electronic	$ 29.00 □	$ 15.00 □
Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of:	Print	$ 36.00 □	$ 18.00 □
	Electronic	$ 6.00 □	$ 3.00 □
	Combo-Print & Electronic	$ 40.00 □	$ 20.00 □
IEEE Sensors Journal:	Print	$ 50.00 □	$ 25.00 □
	Electronic	$ 30.00 □	$ 15.00 □
	Combo-Print & Electronic	$ 63.00 □	$ 32.00 □
***NEW!* Affective Computing, IEEE Transactions on:**	Electronic	$ 30.00 □	$ 15.00 □
***NEW!* Biometrics Compendium**	Online	$ 30.00 □	$ 15.00 □
***NEW!* Smart Grid, IEEE Transactions on**	Print	$ 50.00 □	$ 25.00 □
	Electronic	$ 40.00 □	$ 20.00 □
	Combo-Print & Electronic	$ 60.00 □	$ 30.00 □
IEEE Transactions on Engineering Management	Combo-Print & Electronic	$ 36.00 □	$ 18.00 □
IEEE Engineering Management Review	Combo-Print & Electronic	$ 30.00 □	$ 15.00 □
IEEE Technology Management Package	Combo-Print & Electronic	$ 55.00 □	$ 28.00 □
Wireless Communications, IEEE Transactions on:	Electronic	$ 40.00 □	$ 20.00 □
	Combo-Print & Electronic	$ 85.00 □	$ 43.00 □

*IEEE membership required or requested
Affiliate application to join SP Society only. Amount Paid $________

IEEE Membership Dues $____________

Signal Processing Society Fees $____________

Canadian residents pay 5% GST or 13% HST Reg. No. 125634188 on Society payment(s) and publications only Tax $____________

AMOUNT PAID WITH APPLICATION.................................... TOTAL $____________

Prices subject to change without notice.

□ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
□ **American Express** □ **VISA** □ **MasterCard**
□ **Diners Club** □ **Eurocard**

Exp. Date Mo./Yr.

Cardholder 5 Digit Zip Code Billing Statement Address/USA Only

CALL FOR PROPOSALS

2013, 2014, and 2015

IEEE International Symposium on Biomedical Imaging (ISBI)
Sponsored By
The IEEE Signal Processing Society and
The IEEE Engineering in Medicine and Biology Society

This Call for Proposal is distributed on behalf of the Steering Committee for the IEEE International Symposium on Biomedical Imaging (ISBI) to be held late spring or early summer each year. The IEEE International Symposium on Biomedical Imaging (ISBI) is the premier forum for the presentation of technological advances in theoretical and applied biomedical imaging. ISBI 2013 will be the tenth meeting in this series.

ISBI has played a leading role in facilitating interaction between researchers in medical and biological imaging. The 2013, 2014, and 2015 meetings will continue this tradition of fostering cross-fertilization among different imaging communities and contributing to an integrative approach to biomedical imaging across all scales of observation.

All proposals will be considered. However, the Steering Committee would prefer to host ISBI 2013 in North America.

The conference organizing team is advised to incorporate into their proposal the following items.

- Proposed Dates (late spring or summer)
- Organizing Committee Members
    - Name
    - Biographical information
- Technical Committee Members
    - Name
    - Biographical information
    - Membership in the Bio Imaging And Signal Processing TC (BISP) and/or the Biomedical Imaging And Image Processing TC (BIIP)
- List of bioimaging scientist and research groups who reside in the local area who are in favor of the proposal and who are committed to attend and participate.
- Proposed budget. (For advice on building an IEEE budget please contact Linda Skeahan at l.skeahan@ieee.org.)
- Support that can be anticipated from the local government, universities and or corporations
- Why this location?
    - Airport information
    - Customs and Visa regulations
    - Hotel and convention center information (i.e. space diagrams, maps, etc.)
    - Tourist destinations (i.e. museums, natural wonders, etc.)
    - Average weather conditions for the time of year

Submission of Proposal
Proposals for ISBI are currently being accepted for the 2013, 2014, and 2015.
Proposals for 2013 should be sent no later than **1 July 2010**. Send the proposal to Lisa Schwarzbek, Conference Services Manager, IEEE Signal Processing Society (l.schwarzbek@ieee.org). Notification of acceptance will be made **August 2010**.
Proposals for 2014 and 2015 should be sent no later than **18 March 2011**. Notification of acceptance will be made after ISBI 2011 in Chicago, Illinois.

Proposal Presentation
Proposals that are of interest to the Steering Committee may be asked to present their proposal at a Steering Committee meeting.