# Perception-Based Media Processing

By LINA J. KARAM, *Fellow, IEEE*
*Guest Editor*

W. BASTIAAN KLEIJN, *Fellow, IEEE*
*Guest Editor*

KARON MACLEAN
*Guest Editor*

## I. OVERVIEW

Digital multimedia technologies and services are becoming an inherent part of our daily lives, enabling such broad and commercially significant application domains as entertainment and communications, security and surveillance, education and training, manufacturing, and health, to name a few.

Media processing is ubiquitous in modern society. Activities such as making a mobile or Internet phone call, talking to an automated call-answering system, listening to music, watching television or Internet video, taking or sharing pictures, playing electronic games, and making a diagnosis, all make extensive use of media processing. In addition to applications such as remote sensing, manufacturing, health, security, and surveillance, the use of media processing has also been on the rise in applications such as education and training, as evidenced with the significant increase in the number of electronic books, e-readers, multimedia documents, and multimedia educational Internet content. The use of media processing will continue to expand as existing technologies get refined and as new technologies such as immersive two-way communication, virtual environments, and interactive human–machine interfaces become part of our everyday life. All these media-based applications not only require media processing but they also engage one or more of the human senses (audition, vision, touch) and, therefore, human perception. It follows that the quality/distortion due to processing is best defined as whatever human subjects perceive it to be, and that more effective media-based applications can be developed by designing, implementing, and testing media processing systems that are optimized based on perceptual measures.

This special issue provides a timely review of the state of the art in the areas of perception-based audio, visual, and haptic processing.

Traditionally, multimedia technologies have been developed with a focus on optimizing device- and network-centric measures (such as power, memory, delay, bit rate, error rate, and packet loss rate) rather than user-centric measures. The performance and quality offered by these traditional measures are insufficient to meet the users' increasingly high expectations and needs, especially for emerging interactive multimedia applications. Since the human being is the ultimate consumer of multimedia technologies, it is important to make these technologies more user centric and, as a result, more effective, useful, and enjoyable. For this purpose, there is a need for perception-based multimedia technologies that integrate both technology and human perception/cognition aspects to achieve an enhanced human effectiveness across a broad range of applications. From the human perspective, this will lead to more effective and natural human–machine interfaces and a greatly improved quality of experience (QoE). From the technical perspective, the

Digital Object Identifier: 10.1109/JPROC.2013.2270898

human perception and cognitive aspects can be exploited in designing more effective multimedia systems with an improved performance in terms of reduced computations, latency, power, bit rate, and error rate.

Of the five major human senses of audition, vision, taction (touch), olfaction, and gustation, media-based technologies have mainly targeted audition (speech, audio, music) and vision (image, video, graphics, visualization) and, more recently, haptics (encompassing taction, through active tactile displays, and proprioception, via force feedback interfaces). This Special Issue on Perception-Based Media Processing provides a timely review of the state of the art in the areas of perception-based audio, visual, and haptic processing, and covers key technologies and applications that emphasize perceptual measures as the basis for optimizing the processing of multimedia content.

## II. TOPICS AND PAPERS

The papers in this special issue cover three broad areas: perception-based auditory processing, perception-based visual processing, and perception-based haptic processing.

The first group of papers relates to sound and its perception. First, we discuss technologies that facilitate the transmission, rendering, and retrieval of audio signals and auditory scenes. Each paper highlights the relevance of human perception in this context. One may wonder if in a society of humans, machines are taught to perceive the world as we do. We address this issue with a paper on the similarity of speech recognition in humans and machines. The final paper explains how technology can be used to improve hearing for persons with a hearing impairment.

The second group of papers reviews key technologies and techniques related to vision and its perception. First, we begin with computer graphics technologies for realistic image synthesis. This is followed by methods and technologies for the compression, transmission, quality assessment, and processing of visual data including image, video, and visual textures. Each paper emphasizes the integration of human perception in the design and implementation of the presented methods and technologies. Then, we address the emerging area of attentive visual sensing and processing, which exploits very recent theories of human visual selective attention in various applications.

Finally, the third group surveys representative contributions by leading researchers in the field of haptics, beginning with practical rendering methods in force feedback and taction, backed up by reviews of our current knowledge of tactile perception of both surface properties and the increasingly ubiquitous vibrotactile display. Then, we bring a special focus to the primary emerging application areas for touch feedback in attention management for applications in multitasking environments, with two papers on cueing and modeling of multimodal attention; and close the issue with a model of haptic aesthetic perception.

We present a brief summary of papers in each of these three categories in the following sections.

### A. Auditory

The coding of audio signals is a technology that has become an integral part of modern society. Efficient coding of a signal requires the removal of both irrelevant and redundant information from the signal. Knowledge of human perception is a central aspect as it indicates which attributes of the signal are irrelevant and can be omitted. In the paper "Perceptual coding of high-quality digital audio," Brandenburg *et al.* discuss how irrelevancy and redundancy are removed from audio signals in modern audio coding systems.

In many audio-rendering scenarios, it is desirable to reproduce a recorded or virtual auditory scene. Common multichannel audio formats form a first step in this direction. In recent years, significant progress has been made toward the accurate reproduction of sound fields. However, the interplay between perception and the accuracy of various attributes of the sound field is as yet relatively poorly understood. In the paper "Spatial sound with loudspeakers and its perception: A review of the current state," Spors *et al.* provide a detailed overview of the current state of both the art of spatial sound rendering and knowledge of its perception.

With the explosive increase of media content, particularly in the cloud, audio indexing and classification plays an increasingly important role. Audio indexing identifies and labels acoustic events facilitating quick and efficient access. The retrieval can be based on an audio classification with a fixed set of labels or, in the case of semantic retrieval, on approximate language level descriptions. As the final user is human, the indexing and classification must be perceptually relevant. In the paper "An overview on perceptually motivated audio indexing and classification," Richard *et al.* discuss how perceptual relevance can be achieved. Three basic approaches are used. In the first approach, knowledge of the human auditory system is used to define perceptually relevant features that are then used for characterizing the content. In the second approach, generic features are selected as perceptually relevant by correlations with perceptual test results. In the final approach, users are directly involved to provide feedback in the form of, for example, language-level descriptions.

Well-defined measures of perceived quality of audio and speech are needed for the design and operation of modern telecommunications infrastructure and the design of user devices. Basic measures are based on listening tests involving a set of human listeners. As such tests are expensive and cannot be performed in real time, methods for the objective estimation of speech and audio quality have been developed. Speech intelligibility plays a special role, which often is affected differently by signal distortion than speech quality. In fact, enhancement algorithms that improve perceived quality, often result in decreased intelligibility. In the paper "Objective estimation of speech quality for communication systems," Möller and Heusdens provide an overview of methods for the objective estimation of speech quality and of speech intelligibility. Their paper discusses the perceptual and cognitive bases for quality prediction, the basic approach to the prediction of overall quality, the prediction of specific attributes of quality, one of which is intelligibility, as well as some example applications.

Humans interpret speech seemingly without effort. It would be natural to base the development of automatic speech recognition on exploiting knowledge of the auditory systems that determine human perception of speech. In practice, automatic speech recognition has advanced through a sequence of engineering solutions that sometimes were inspired by the auditory example, and sometimes not. In the paper "Perceptual properties of current speech recognition technology," Hermansky *et al.* discuss how automatic speech recognizers have been made robust to the variability in timbre and speaking rate between talkers, how they have been made robust to distortion and additive noise, and how such systems can exploit spatial diversity. For each of these topics, they then compare the technical approach with knowledge of the human auditory system. Their conclusion is that modern speech recognition technology is largely consistent with attributes of the auditory system.

Technology now allows a significant improvement in the perception of sound for persons with hearing impairments. People with profound impairments often receive cochlear implants that directly stimulate the auditory nerve, enabling them to partake in a society set up for normal-hearing persons. Hearing aids significantly improve sound perception for persons with a mild or moderate hearing loss. In the paper "Sound processing for better coding of monaural and binaural cues in auditory prostheses," Wouters *et al.* discuss in detail the signal processing advances that are central to the rapidly improving performance of hearing aids and cochlear implants.

### B. Visual

Reproducing a realistic visual scene is a computer graphics technology that is central to many applications from movie production, advertisement, modeling and design, manufacturing, virtual environments and training, to name a few. In the paper "On visual realism of synthesized imagery," Reinhard *et al.* discuss how human perception is exploited for the design of visual modeling and rendering systems with improved visual realism at lower computational complexity.

Many factors affect or impair the quality of visual content, including acquisition, processing, compression, transmission, protection, display, and reproduction systems. Finding effective ways to monitor and control the perceived quality of visual media is key to enabling many emerging applications. In the paper "Automatic prediction of perceptual image and video quality," Bovik discusses the principles and methods of modern algorithms for automatically assessing the perceived quality of visual signals.

As the volume of visual content being transported and viewed continues to increase exponentially and is predicted to dominate the mobile traffic in the very near future, the compression of the visual content for efficient transmission and storage has become a pressing need. As for audio coding, visual coding relies on the removal of redundant and less relevant components. The modeling and integration of the human visual perception in visual compression systems is key to omit perceptually redundant and less relevant signal components. In the paper "Perceptual visual signal compression and transmission," Wu *et al.* discuss the principles behind perception-based visual compression, and they provide an overview of visual coding systems that integrate human perception aspects for optimizing the perceived quality or for optimizing the needed bit rate (bandwidth) while achieving a desired perceived visual quality.

The accumulation of large collections of digital visual content has created the need for efficient and intelligent schemes for image analysis, classification, and indexing for content-based image retrieval. Since humans are the ultimate users of most image retrieval systems, it is important to organize the content semantically, according to meaningful categories. This requires an understanding of the important semantic categories that humans use for image classification, and the extraction of meaningful perceptual-based image features that can discriminate between these categories. In particular, knowledge of how human perceive visual textures is central to these and other applications. In the paper "Image analysis: Focus on texture similarity," Pappas *et al.* discuss key aspects in visual texture perception and how these are used in the development of perception-based texture similarity metrics with applications in visual compression and content-based retrieval.

Humans are surrounded by complex environments. The human brain has, however, limited resources and uses these resources in a selective way to capture the most relevant information to be processed for guidance, survival, and decision making in a given situation. This is known as selective attention. The recent emerging area of attentive media sensing and processing exploits human attention models for capturing and processing at higher fidelity information at attended locations, and for discarding or processing at a lower fidelity insignificant parts of the multimedia content. In the paper "Visual attention and applications in multimedia technologies," Le Callet and Niebur discuss human visual selective attention and challenges related to modeling visual attention. They also provide an overview of applications and technologies that integrate visual attention models.

### C. Haptic

Rendering touchable environments using force feedback was the first haptic technology to emerge in the early 1990s: these systems impose active forces and motions on the user, which are controlled in a similar manner to a robotic device and sensed through the user's proprioception. Since then, algorithms and hardware have evolved in lockstep with our perceptual research that has provided performance specifications. In the paper "Representations and algorithms for force-feedback display," Otaduy *et al.* provide an update on today's force rendering algorithms, typically used to present interactive 3-D environments in tandem with closely coupled visual and auditory stimuli at high bandwidths, with a focus on computational aspects of collision detection, dynamics simulation, and constrained optimization.

Scientific study of tactile perception—what we feel through our skin—certainly predates haptic display technology, but has advanced rapidly with access to controllable test platforms. In the paper "Haptic perception of material properties and implications for applications,'' Klatzky *et al.* observe that the touch sense is better at assessing material properties than shape, and review current knowledge of both material perception and the wide diversity of algorithms and devices that are used to render them in synthetic environments, including roughness, friction, and thermal properties. In the paper "Vibrotactile display: Perception, technology, and applications," Choi and Kuchenbecker spotlight the vibrotactile display, which activates a different set of mechanoreceptors and can be accessed through very low-cost and easily embedded, but equally easily misused, vibrotactile displays.

Haptics has vast potential as a vehicle for reducing the load on our attention, which is increasingly beseiged as our computational technology becomes ever more situated—handheld, embedded, or otherwise—in the world around us. The touch sense is an alternative to vision and audition but also has the capacity to exacerbate sensory overload if used unwisely. In the paper "Multimodal support for interruption management: Models, empirical findings, and design recommendations," Sarter outlines model-based approaches to distributing multiodal information across sensory channels to achieve desired performance in detection, interpretation, and appropriate handling of interrupting tasks and signals. In the paper "Efficient multimodal cuing of spatial attention," Gray *et al.* detail how this can work in the important example of multimodal spatial cueing, and ground it in several classes of applications.

Aesthetic perception is of growing importance as designed haptic properties enter the consumer world—how do we measure and predict how people will respond emotionally to the things they feel, and how does this response develop? In the paper "A model for haptic aesthetic processing and its implications for design," Carbon and Jakesch develop a functional, stage-based model of haptic aesthetic processing and relate it to real-world design issues.

We hope that the readers of the Proceedings of the IEEE will enjoy the timely collection of articles in this issue. We would like to thank all the authors and reviewers for their invaluable contributions and efforts. We would also like to thank the IEEE Managing Editor Jim Calder for his support of the issue, and the Proceedings of the IEEE staff, Jo Sun and Margery Meyer, for their assistance with the preparation of this issue. ■

## ABOUT THE GUEST EDITORS

**Lina J. Karam** (Fellow, IEEE) received the B.E. degree in computer and communications engineering from the American University of Beirut, Beirut, Lebanon, in 1989 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1992 and 1995, respectively.

She is a Full Professor in the School of Electrical, Computer & Energy Engineering, Arizona State University, Phoenix, AZ, USA, where she directs the Image, Video, & Usabilty (IVU) Research Laboratory. Her industrial experience includes image and video compression development at AT&T Bell Labs, Murray Hill, NJ, USA, multidimensional data processing and visualization at Schlumberger, and collaboration on computer vision, image/video processing, compression, and transmission projects with industries including Intel, NTT, Motorola/Freescale, General Dynamics, and NASA. She has over 100 technical publications and she is a coinventor on a number of patents.

Dr. Karam was awarded a U.S. National Science Foundation CAREER Award, a NASA Technical Innovation Award, and the 2012 Intel Outstanding Researcher Award. She was also awarded the Outstanding Faculty Award by the IEEE Phoenix Section in 2012. She has served on several journal editorial boards, several conference organization committees, and several IEEE technical committees. She served as the Technical Program Chair of the 2009 IEEE International Conference on Image Processing, as the General Chair of the 2011 IEEE International DSP/SPE Workshops, and as the Lead Guest Editor of the IEEE Journal on Selected Topics in Signal Processing, Special Issue on Visual Quality Assessment. She has cofounded two international workshops (VPQM and QoMEX). She is currently serving as the General Chair of the 2016 IEEE International Conference on Image Processing. She is currently a member of the IEEE Signal Processing Society's Multimedia Signal Processing Technical Committee and the IEEE Circuits and Systems Society's DSP Technical Committee. She is a member of the Signal Processing, Circuits and Systems, and Communications societies of the IEEE.

**W. Bastiaan Kleijn** (Fellow, IEEE) received the M.S. degree in physics and the Ph.D. degree in soil science from the University of California, Riverside, Riverside, CA, USA, both in 1981, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1984, and the Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 1991.

He has been a Professor at Victoria University of Wellington, Wellington, New Zealand, since 2010. He is also a Professor at Delft University of Technology (part-time, since 2011) and at the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was the Head of the Sound and Image Laboratory. Before joining KTH in 1996, he worked at AT&T Bell Laboratories (Research) on speech processing. He was a founder of Global IP Solutions, which developed voice and video processing technology, and was acquired by Google in 2010. He has authored or coauthored over 230 peer-reviewed papers and holds close to 40 U.S. patents.

Prof. Kleijn has served or is serving on the Editorial Boards of IEEE Signal Processing Letters, the IEEE Transactions on Speech and Audio, IEEE Signal Processing Magazine, and *Signal Processing*. He was Technical Chair of the 1999 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the 2011 European Signal Processing Conference (EUSIPCO).

**Karon MacLean** received the B.Sc. degree in biology and mechanical engineering from Stanford University, Stanford, CA, USA, in 1986 and the M.Sc. and Ph.D. degrees in mechanical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1996, with professional robotics engineering (Center for Engineering Design, University of Utah) in between.

She is a Professor in Computer Science at the University of British Columbia (UBC), Vancouver, BC, Canada. She worked as a Research Scientist at Interval Research, Palo Alto, CA, USA, coming to UBC in 2000. Her research in ubiquitous haptic and multimodal interfaces brings together robotics, interaction, and affect design and psychology with the larger goal of restoring physicality to embedded computation, and has been recently supported by General Motors, Nokia, Immersion, Nissan, and others. She uses touch feedback as part of a multisensory HCI toolbox in the context of real design problems like mobile devices and automobile controls, to leverage new design techniques and define her studies of multimodal perception and attention.

Prof. MacLean was the 2001 Peter Wall Early Career Scholar and received the 2007 Izzak Walton Killam Memorial Faculty Research Fellowship and the 2008 Charles A. McDowell Award. She is an Associate Editor of the IEEE Transactions on Haptics, founding member of several other editorial and advisory boards, and Co-Chair of the 2010 and 2012 IEEE Haptics Symposium.