Multimed Tools Appl (2016) 75:5047–5053
DOI 10.1007/s11042-016-3443-x

GUEST EDITORIAL

Guest Editorial: Immersive Audio/Visual Systems

Lei Xie[1] · Longbiao Wang[2] · Janne Heikkilä[3] · Peng Zhang[1]

Published online: 22 April 2016

1 Introduction

With recent advances in audio/visual information processing and various sensory technologies, we are able to collect rich data, e.g., facial expression, eye gaze, head movement, body gesture, expressive speech and audio scene, from multiple sensors and diverse devices. We are on our fast way to realize immersive user experiences on auditory and visual spaces with these collected rich data. We envision that in the future an immersive audio/video system can enable natural and immersive experiences for users via advanced audio/visual technologies. In recent years, many researchers have been studying on audio/visual processing for immersive environments and human-computer interaction, including media data (audio, speech, video) capture, synthesis, enhancement and recognition, 3D audio/visual rendering and construction, and multimodal interaction. For example, through advanced audio and visual technologies, the fast emerging augmented reality (AR) systems provide users immersive experiences that have never been seen before.

✉ Lei Xie
lxie@nwpu.edu.cn

Longbiao Wang
wang@vos.nagaokaut.ac.jp

Janne Heikkilä
jth@ee.oulu.fi

Peng Zhang
zh0036ng@nwpu.edu.cn

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[2] Nagaoka University of Technology, Nagaoka, Japan

[3] Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland

The goal of this special issue is to bring together researchers and technologists engaged in the development of immersive audio/visual systems. We received about 30 high-quality submissions and each paper was strictly peer-reviewed. After the first and second rounds of review, 17 manuscripts were finally selected to be included in this special issue.

2 Recognizing humans and understanding their behaviors

Recognizing humans and understanding their behaviors are indispensable stages in creating immersive user experiences. In this special issue, Zheng et al. [16] focus on recognizing human faces with low dimensional feature representation. They implant the traditional neural network, multi-layer perceptrons (MLP), in a siamese architecture to realize flexible dimensionality reduction but maintain good recognition performance. Another important task in advanced large-scale immersive AR environments is the long-term object tracking to feel the changes among the synthetic stereoscopic image sequences. To achieve this purpose, Zhang et al. [15] propose a novel Bayesian tracking fusion framework with online classifier ensemble strategy. The tracking formulates a fusion framework for online learning of multiple trackers by modeling a cumulative loss minimization process. With an optimal pair-wise sampling scheme for the SVM classifier, the proposed fusion framework can achieve more accurate tracking performance for more applicable immersive AR systems.

As we know, speech recognition has been providing an immersive communication means. However, in a voice-activated immersive scenario like smart-home, we desire to talk from distance to the smart device.[1] Recognizing human speech from distance, known as distant or far-field speech recognition, is a challenge task because of complicated environmental noises and undesirable room reverberations. In this special issue, Ren et al. [10] propose three integration schemes for robust distant-talking speech recognition which combine bottleneck feature extraction with dereverberation technique. As an accompanying paper by the same institution, Phapatanaburi et al. [9] propose a combination of Gaussian Mixture Models (GMM) and Deep Neural Networks (DNNs) to identify the speaker accent in reverberant environments.

When talking, we often move heads and exhibit various facial expressions. Non-verbal cues, e.g., hand gestures and head motions, are used to express feelings, give feedbacks and engage immersive human-human communication. In [14], Yang et al. study the relations between speech and head motion and learn a bimodal mapping from speech to head motions. Specifically, they give an interesting investigation to discover what kinds of prosodic and linguistic features have the most significant influence on emotional head motions.

3 Immersive sound and visual display

Immersive sound and visual display is another popular topic. In this special issue, Fang and Wang [3] address the problem of reproducing a scene with high contrast ratios when using projection-based immersive visual systems [2]. A lack of sufficient dynamic range deteriorates the user experience, which has provided the motivation for a novel system where the

[1] For example, Amazon Echo is a wireless speaker and voice command device from Amazon.com.

projector is coupled with a spot light that generates much higher illuminance to the screen than the surrounding saturated area. A galvanometer is used to deflect the beam which enables tracking the movement of the bright light sources displayed on the screen.

There are three papers focusing on immersive audio and speech rendering. Virtual acoustic space (VAS), also known as virtual auditory space, is a technique in which sounds presented over headphones appear to originate from any desired direction in space. Kang et al. [7] propose a method to synthesize virtual sound sources near the listener position due to the differences between sound fields of real and virtual sound sources. Using a rigid sphere as a model of humans head within 1 m, the authors demonstrate how the Interaural Level Difference (ILD) is influenced by the reproduction artifacts of a line array. There are regions of virtual source locations near the listener within which correct ILDs can be provided. Zheng et al. [17] describe a system for encoding and communicating navigable speech soundfields. A low-delay Direction of Arrival (DOA)-based frequency domain sound source separation approach is proposed that requires only 250 ms of speech signal. Joint compression is achieved through a previously proposed perceptual analysis-by-synthesis spatial audio coding scheme that encodes sources into a mixture signal that can be compressed by a standard speech codec at 32 kbps.

Room impulse response (RIR) is widely used in immersive audio communication, which contributes to the illusion of a virtual sound source outside the listener's head. However, due to the underlying complexity of sound propagation, the acoustic model of RIR is very computation-intensive. To solve this problem, Fu et al. [4] propose an approach that uses graphics processing units (GPU) to greatly speed-up the calculation.

4 The importance of speech

The speech science is the theoretical foundation for immersive audio and speech systems. In this special issue, there are two papers focusing on articulatory phonetics and speech production. Wei et al. [11] propose a novel deep learning framework for the creation of a bidirectional mapping between articulatory information and synchronized speech recorded using an ultrasound system. Constructing a mapping between articulatory movements and corresponding speech could significantly facilitate speech training and the development of speech aids for voice disorder patients. In another paper [12], an ultrasound system is used in combination with the electromagnetic articulography (EMA) system to record the multi-modality movement of the tongue. The EMA and ultrasound data were registered and matched to the same audio signal, after which these two sets of data were fused for each time point. In addition, a method for vocal tract shape reconstruction and modeling is proposed for the ultrasound dataset by using an active shape model.

The unique speech production system of each individual creates personalized voice. In an immersive application, we expect to generate an arbitrary speaker's voice with a small size of data, creating personal voices for all. Voice conversion, modifying the recorded speech of a source speaker toward a given target speaker, is a popular way to achieve such voice personalization. In this special issue, Nguyen et al. [8] provide a comprehensive voice conversion framework using deep neural networks to convert both timbre and prosodic features. Experiments show that the use of prosodic and high-resolution spectral features leads to high-quality converted speech.

Visual speech, i.e., speech-evoked facial movements, plays an indispensable role in speech communication. Therefore, human-machine speech commutation will become more

immersive if a vivid talking head is present. In [2], Fan et al. propose a deep bidirectional long short-term memory (DBLSTM) approach in modeling the long contextual, nonlinear mapping between audio and visual streams for video-realistic talking head. Their study shows that the proposed DBLSTM approach outperforms the hidden Markov model (HMM) approach in both objective and subjective evaluations.

5 Security of immersive audio/visual systems

Besides the audio/visual techniques widely used in various types of immersive systems, which have been discussed in this special issue, some researchers have pointed out the importance of security of such systems. Wu et al. [13] conduct a systematic analysis of text-dependent speaker verification systems to speech replay and voice conversion attacks. Specifically, using the same protocol and database, they analyze the interplay of voice conversion and speaker verification by linking the voice conversion objective evaluation measures with the speaker verification error rates to investigate the vulnerabilities from the perspective of voice conversion.

Multimodal information is usually used to improve the system robustness. For example, audiovisual systems equipped with a microphone and a front-facing video camera enable biometric person verification based on face and voice. However, such systems are often vulnerable to spoofing attacks where an unauthorized user tries to access the system by falsifying the audiovisual biometric data. To deal with this problem Boutellaa et al. [1] present a new spatiotemporal method for talking face verification by employing speech synchrony detection to ensure the liveness of a subject.

Immersive audio/visual systems often employ wireless communication techniques for data transmission. When private data is transmitted wirelessly between legitimate users it creates a potential risk for malicious attacks such as passive eavesdropping. In order to improve the security of video transmission, Hussain et al. [6] propose a new a physical layer security method, termed noise aggregation, for limiting the amount of information that can be eavesdropped by an unauthorized user.

6 Immersive tools

Multimedia-based immersive tools also play increasingly important roles in many multidisciplinary applications, such as clinical diagnosis, at the current stage. Huang et al. [5] propose a novel medical imaging-based immersive diagnosis tool for the dementia disease, which is a serious non-communicable disease in many aging societies. Technically, a new surrogate pair-wise ranking evaluation measure is proposed based on the conventional Kendall-Tau correlation coefficient, and a parameterized similarity measure is determined automatically based on the new evaluation measure via a learning strategy. This new immersive tool is statistically proved to be promising based on a large number of real-life database composed of 350 demented patients.

7 Summary

The special issue covers a wide range of methods and technologies towards immersive audio/visual systems, namely recognizing humans and understanding their behaviors,

immersive sound and visual display, the importance of speech, Security of immersive audio/visual systems and immersive tools. We hope that the readers will find these papers informative and interesting. We would like to thank the authors of all submitted papers. We also wish to offer our sincere thanks to the Editor-in-Chief, Professor Borko Furht and to all editorial staffs for their valuable supports throughout the preparation and publication of this special issue. We also thank to the reviewers for their help in reviewing the papers. We believe that, with the fast development of virtual/augmented reality and internet of things, immersive audio/visual systems will play a decisive role in the near future.

References

1. Boutellaa E, Boulkenafet Z, Komulainen J, Hadid A (2015) Audiovisual synchrony assessment for replay attack detection in talking face biometrics. Multimed Tools Appl. doi:10.1007/s11042-015-2848-2
2. Fan B, Xie L, Yang S, Wang L, Soong FK (2015) A deep bidirectional lstm approach for video-realistic talking head. Multimed Tools Appl. doi:10.1007/s11042-015-2944-3
3. Fang C, Wang Y (2015) Light source imitation by using galvanometer scanner and spot light. Multimed Tools Appl. doi:10.1007/s11042-015-2942-5
4. Fu Z-H, Li J-W (2015) Gpu-based image method for room impulse response calculation. Multimed Tools Appl. doi:10.1007/s11042-015-2943-4
5. Huang W, Zeng S, Li J, Chen G (2015) A new image-based immersive tool for dementia diagnosis using pairwise ranking and learning. Multimed Tools Appl. doi:10.1007/s11042-015-2826-8
6. Hussain M, Du Q, Sun L, Ren P (2015) Security enhancement for video transmission via noise aggregation in immersive systems. Multimed Tools Appl. doi:10.1007/s11042-015-2936-3
7. Kang D-S, Choi J-W, Martens WL (2015) Distance perception of a virtual sound source synthesized near the listener position. Multimed Tools Appl. doi:10.1007/s11042-015-2878-9
8. Nguyen HQ, Lee SW, Tian X, Dong M, Chng ES (2015) High quality voice conversion using prosodic and high-resolution spectral features. Multimed Tools Appl. doi:10.1007/s11042-015-3039-x
9. Phapatanaburi K, Wang L, Sakagami R, Zhang Z, Li X, Iwahashi M (2015) Distant-talking accent recognition by combining gmm and dnn. Multimed Tools Appl. doi:10.1007/s11042-015-2935-4
10. Ren B, Wang L, Lu L, Ueda Y, Kai A (2015) Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition. Multimed Tools Appl. doi:10.1007/s11042-015-2849-1
11. Wei J, Fang Q, Zheng X, Lu W, He Y, Dang J (2015) Mapping ultrasound-based articulatory images and vowel sounds with deep neural network framework. Multimed Tools Appl. doi:10.1007/s11042-015-3038-y
12. Wei J, Wang S, Lu W, Hou Q, Fang Q, Dang J (2015) Multi-modal recording and modeling of vocal tract movements. Multimed Tools Appl. doi:10.1007/s11042-015-3040-4
13. Wu Z, Li H (2015) On the study of replay and voice conversion attacks to text-dependent speaker verification. Multimed Tools Appl. doi:10.1007/s11042-015-3080-9
14. Yang M, Jiang J, Tao J, Mu K, Li H (2016) Emotional head motion predicting from prosodic and linguistic features. Multimed Tools Appl. doi:10.1007/s11042-016-3405-3
15. Zhang P, Zhuo T, Zhang Y, Huang H, Chen K (2015) Bayesian tracking fusion framework with online classifier ensemble for immersive visual applications. Multimed Tools Appl. doi:10.1007/s11042-015-2827-7
16. Zheng L, Duffner S, Idrissi K, Garcia C, Baskurt A (2015) Siamese multi-layer perceptrons for dimensionality reduction and face identification. Multimed Tools Appl. doi:10.1007/s11042-015-2847-3
17. Zheng X, Ritz C, Xi J (2015) Encoding and communicating navigable speech soundfields. Multimed Tools Appl. doi:10.1007/s11042-015-2989-3

Lei Xie received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2004. He is currently a Professor with School of Computer Science, Northwestern Polytechnical University, Xi'an, China. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium, as a Visiting Scientist. From 2004 to 2006, he was a Senior Research Associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Hong Kong. From 2006 to 2007, he was a Postdoctoral Fellow in the Human-Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. He has published more than 120 papers in major journals and conference proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, INFORMATION SCIENCES, PATTERN RECOGNITION, ACL, ACM Multimedia, Interspeech, ICPR, ICME and ICASSP. He has served as program chairs and organizing chairs in various conferences. He is a Senior Member of IEEE. His current research interests include speech and language processing, multimedia and human-computer interaction.

Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. and Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2005 and 2008 respectively. From July 2000 to August 2002, he worked at the China Construction Bank. He was an assistant professor in the faculty of Engineering at Shizuoka University, Japan from April 2008 to September 2012. Since October 2012 he has been an associate professor at Nagaoka University of Technology, Japan. His research interests include robust speech recognition, speaker recognition and acoustic signal processing. He received the "Chinese Government Award for Outstanding Self-financed Students Abroad" in 2008. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Janne Heikkilä received his Doctor of Science in Technology degree in Information Engineering from the University of Oulu in 1998. Currently, he is a professor of computer vision and digital video processing in the Faculty of Information Technology and Electrical Engineering at the University of Oulu, and the head of the Degree Program in Computer Science and Engineering. He has served as an area chair and a member of program and organizing committees of several international conferences. He is an associate editor of IET Computer Vision, Journal of Electronic Imaging and Electronic Letters on Computer Vision and Image Processing, a member of the governing board of the International Association for Pattern Recognition (IAPR), and a senior member of the IEEE. During 2006-2009 he was the president of the Pattern Recognition Society of Finland. He has been the principal investigator in numerous research projects funded by Academy of Finland, National Agency for Technology and Innovation (Tekes) and enterprises. His research interests include computer vision, machine learning, digital image and video processing, and biomedical image analysis. He has supervised 9 completed doctoral dissertations and published more than 160 peer reviewed scientific articles in international journals and conferences.

Peng Zhang received the B.E. degree from the Xian Jiaotong University, China in 2001. He received his PhD from Nanyang Technological University, Singapore in 2011. He is now an associate professor in School of Computer Science, Northwestern Polytechnical University, China. His current research interests include signal processing, multimedia security and pattern recognition. He acts as the technical committee and the reviewer in several important international conferences and journals including ICIP, ICME, IEEE T-IP, IEEE T-CSVT, MTAP and Neurocomputing. He is a member of ACM.