Multimodal joint information processing in human machine interaction: recent advances

Lei Xie · Zhigang Deng · Stephen Cox

Published online: 5 November 2013
© Springer Science+Business Media New York 2013

Interaction between humans is a multimodal process in nature, which integrates the different modalities of vision, hearing, gesture and touch. It is not surprising then that conventional uni-modal human-machine interactions lag in performance, robustness and naturalness when compared with human-human interactions. Recently, there has been increasing research interest in jointly processing information in multiple modalities and mimicking human-human multimodal interactions [2, 4, 5, 9, 13, 14, 16, 18, 19, 21, 22]. For example, human speech production and perception are bimodal in nature: visual cues have a broad influence on perceived auditory stimuli [17]. The latest research in speech processing has shown that integration of both auditory and visual information, i.e., acoustic speech combined with facial and lip motions, achieves significant performance improvement in tasks such as speech recognition [3] and emotion recognition [16]. However, information from difference modalities can be difficult to integrate: the representations of different signals are heterogeneous and the signals are often asynchronous and loosely-coupled. Therefore, finding more effective ways to integrate and jointly process information from different modalities is essential for the success of such multimodal human-machine interactive systems. Meanwhile, multimodal applications are becoming increasingly important, especially for mobile computing and digital entertainment scenarios, such as natural user interfaces (NUI) on smartphones and motion sensing gaming.

This special issue aims to bring together work by researchers and technologists engaged in the development of multimodal technologies for information processing, emerging multimedia applications and user-centric human computer interaction. We received more than

L. Xie (✉)

School of Computer Science, Northwestern Polytechnical University, Xi'an, China
e-mail: lxie@nwpu.edu.cn

Z. Deng

Computer Science Department, University of Houston, Houston, USA
e-mail: zdeng@cs.uh.edu

S. Cox

School of Computing Sciences, University of East Anglia, Norwich, UK
e-mail: S.J.Cox@uea.ac.uk

30 high-quality submissions and each manuscript was peer-reviewed by at least three reviewers. After the first and second rounds of review, 14 manuscripts were finally selected to be included.

This special issue covers a wide range of topics in human-machine interaction, including text- or speech-driven facial animation [11, 23, 28], emphatic speech synthesis [20], head and facial gesture synthesis [10], human pose estimation [24], crowd counting [6], person identification [25] and facial expression recognition [8].

Recognizing humans and understanding their behaviors from multiple modalities are indispensable steps for creating systems that interact naturally with human users. In this special issue, Fu et al. [6] aim to achieve reliable and real-time human counting. They propose a scene-adaptive, accurate and fast crowd counting approach using joint depth and color information. Tsai et al. [25] address the problem of long-range person identification with a multimodal information fusion approach, which includes multiview face detection, height measurement and face recognition. Recognizing human facial expressions is critical to NUI applications that are required to understand human emotions. As a means to this end, Hsu et al. [8] propose an effective facial expression recognition approach using bag of distances. In another paper, Sun et al. [24] deal with the problem of 3D human pose estimation. Specifically, they address the difficulty of high-dimensionality of human pose space and the multimodality of the distribution by a novel motionlet LLC coding approach in a discriminative framework. Last, but not least, data is extremely important in order to understand human behavior, cognition and emotion. In this special issue, Gamboa et al. [7] introduce a new multimodal database, namely HiMotion, which includes both human-computer interaction and psycho-physiological data, collected through experiments with synchronized recordings of keyboard, mouse, and central/peripheral nervous system measurements.

Another goal for NUI is to provide machines with synthesized faces that look, talk and behave like real human faces, which gives human-machine interfaces the feel of immersive human-human interaction. To this end, machines need to synthesize both auditory and visual human speech, i.e., to animate a speaking virtual character with vivid lifelike appearance and expressive synthetic speech [27]. In this special issue, Xie et al. [28] propose a statistical parametric approach to video-realistic text-driven talking avatar. In the approach, multimodal hidden Markov models (HMMs) are learned from audio-visual recordings of a talking subject and a video-realistic talking avatar is synthesized using the HMM trajectory generation algorithm. In an accompanying paper by the same institution, Jiang et al. [11] present a speech-driven facial animation approach that models the audio-visual articulatory process by a specifically-designed dynamic Bayesian network (DBN). In the mobile computing age, talking avatars, or virtual assistants, are playing an increasingly important role in serving a natural user interface due to the physical limits of smartphones and special features of mobile platforms. In [23], Shih et al. aim to synthesize a real-time speech-driven talking face for mobile multimedia. A lifelike talking avatar requires not only natural speech articulation, but also expressive head motions, emotional facial expressions and other meaningful facial gestures [15]. Jia et al. [10] provide an interesting piece of work that synthesizes expressive head and facial gestures using the three dimensional pleasure-displeasure, arousal-nonarousal and dominance-submissiveness (PAD) descriptors of semantic expressivity. Synthesizing human speech has many potential applications. For example, in this special issue, Meng et al. [20] introduce a hidden Markov model (HMM)-based emphatic speech synthesis approach which is used to generate corrective feedback in a computer-aided pronunciation training (CAPT) application.

The other four papers of this special issue focus on diverse research areas related to human-machine interaction. Alcoverro et al. [1] propose an interesting gesture-based

interface designed to interact with panoramic scenes. Wang et al. [26] present an affection-arousal-based highlight extraction approach for soccer video retrieval and summarization. Zhang et al. [29] introduce a web video thumbnail recommendation system with content-aware analysis and query-sensitive matching. Khan et al. [12] focus on a basic problem of image processing. Specifically, they provide a robust scheme for random valued impulsive noise reduction along with edge preservation by anisotropic diffusion with improved diffusivity.

We hope that the readers will find this special issue informative and interesting. We would like to thank the authors of all submitted papers. Unfortunately, we could not include all submissions due to the page limit of the special issue. 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 excellent help in reviewing the manuscripts.

References

- Alcoverro M, Suau X, Morros JR, López-Méndez A, Gil A, Ruiz-Hidalgo J, Casas JR (2013) Gesture control interface for immersive panoramic displays. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1605-7](https://doi.org/10.1007/s11042-013-1605-7)
- Alepis E, Virvou M (2012) Multimodal object oriented user interfaces in mobile affective interaction. *Multimed Tool Appl* 59(1):41–63
- Chen T (2001) Audiovisual speech processing: lip reading and lip synchronization. *IEEE Signal Proc Mag* 18(1):9–21
- Debevc M, Kosec P, Holzinger A (2011) Improving multimodal web accessibility for deaf people: sign language interpreter module. *Multimed Tool Appl* 54(1):181–199
- Ekenel HK, Semela T (2013) Multimodal genre classification of TV programs and youtube videos. *Multimed Tool Appl* 63(2):547–567
- Fu H, Ma H, Xiao H (2013) Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1608-4](https://doi.org/10.1007/s11042-013-1608-4)
- Gamboa H, Silva H, Fred A (2013) Himotion: a new research resource for the study of behavior, cognition, and emotion. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1602-x](https://doi.org/10.1007/s11042-013-1602-x)
- Hsu F-S, Lin W-Y, Tsai T-W Facial expression recognition using bag of distances. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1616-4](https://doi.org/10.1007/s11042-013-1616-4)
- Huang Q, Cox S (2010) Inferring the structure of a tennis game using audio information. *IEEE Trans Audio Speech Lang Process* 19(7):1925–1937
- Jia J, Wu Z, Zhang S, Meng HM, Cai L (2013) Head and facial gestures synthesis using PAD model for an expressive talking avatar. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1604-8](https://doi.org/10.1007/s11042-013-1604-8)
- Jiang D, Zhao Y, Sahli H, Zhang Y (2013) Speech driven photo realistic facial animation based on an articulatory DBN model and AAM features. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1610-x](https://doi.org/10.1007/s11042-013-1610-x)
- Khan NU, Arya KV, Pattanaik M (2013) Edge preservation of impulse noise filtered images by improved anisotropic diffusion. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1620-8](https://doi.org/10.1007/s11042-013-1620-8)
- Khoury EE, Sénéac C, Joly P (2013) Audiovisual diarization of people in video content. *Multimed Tool Appl*. doi:[10.1007/s11042-012-1080-6](https://doi.org/10.1007/s11042-012-1080-6)
- Kijak E, Gravier G, Oisel L, Gros P (2006) Audiovisual integration for tennis broadcast structuring. *Multimed Tool Appl* 30(3):289–311
- Le BH, Ma X, Deng Z (2012) Live speech driven head-and-eye motion generators. *IEEE Trans Vis Comput Graph* 18(11):1902–1914
- Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. *Multimed Tool Appl* 49(2):277–297
- McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746–748
- Mekhaldi D, Lalanne D, Ingold R (2012) A multimodal alignment framework for spoken documents. *Multimed Tool Appl* 61(2):353–388
- Meng H, Oviatt S, Patamianos G (2009) Introduction to the special issue on multimodal processing in speech-based interactions. *IEEE Trans Audio Speech Lang Process* 17(3):409–410

20. Meng F, Wu Z, Jia J, Meng H, Cai L (2013) Synthesizing english emphatic speech for multimodal corrective feedback in computer-aided pronunciation training. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1601-y](https://doi.org/10.1007/s11042-013-1601-y)
21. Montagnuolo M, Messina A (2009) Parallel neural networks for multimodal video genre classification. *Multimed Tool Appl* 41(1):125–159
22. Snoek CG, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. *Multimed Tool Appl* 25(1):5–35
23. Shih P-Y, Paul A, Wang J-F, Chen Y-H (2013) Speech-driven talking face using embedded confusable system for real time mobile multimedia. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1609-3](https://doi.org/10.1007/s11042-013-1609-3)
24. Sun L, Song M, Tao D, Bu J, Chen C (2013) Motionlet LLC coding for discriminative human pose estimation. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1617-3](https://doi.org/10.1007/s11042-013-1617-3)
25. Tsai H-C, Chen B-W, Wang J-F, Paul A (2013) Enhanced long-range personal identification based on multimodal information of human features. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1606-6](https://doi.org/10.1007/s11042-013-1606-6)
26. Wang Z, Yu J, He Y, Guan T (2013) Affection arousal based highlight extraction for soccer video. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1619-1](https://doi.org/10.1007/s11042-013-1619-1)
27. Xie L, Liu Z-Q (2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Trans Multimed* 9(23):500–510
28. Xie L, Sun N, Fan B (2013) A statistical parametric approach to video-realistic text-driven talking avatar. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1633-3](https://doi.org/10.1007/s11042-013-1633-3)
29. Zhang W, Liu C, Wang Z, Li G, Huang Q, Gao W (2013) Web video thumbnail recommendation with content-aware analysis and query-sensitive matching. *Multimed Tool Appl*. doi:[10.1007/s11042-013-1607-5](https://doi.org/10.1007/s11042-013-1607-5)

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 80 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*, *ACM/Springer Multimedia Systems*, *ACL*, *Interspeech*,

ICPR, *ICME* and *ICASSP*. He has served as program chair and organizing chair in several conferences. He is a Senior Member of IEEE, a member of ISCA, a member of ACM, a member of APSIPA and a senior member of China Computer Federation (CCF). He is a Board-of-Governor of the Chinese Information Processing Society of China (CIPSC), a board member of the APSIPA Speech, Language and Audio (SLA) technical committee, a board member of the multimedia technical committee of CCF, a board member of the multimedia technical committee of China Society of Image and Graphics (CSIG). His current research interests include speech and language processing, multimedia and human-computer interaction.

Zhigang Deng is currently an Associate Professor of computer science at the University of Houston (UH) and the founding director of the UH Computer Graphics and Interactive Media (CGIM) Lab. He received the BS degree in mathematics from Xiamen University, China, the MS degree in computer science from Peking University, China, and the PhD degree in computer science from the Department of Computer Science, University of Southern California in 2006. His research interests include computer graphics, computer animation, virtual human modeling and animation, and human computer interaction. He is a senior member of the IEEE, a member of the ACM, and a Founding Board member of the International Chinese Association of Human Computer Interaction. Besides serving as the Associate Editor of the *Computer Animation and Virtual Worlds Journal* (Wiley) and the *ACM Computer In Entertainment* magazine, he is also the Conference Co-Chair of the IEEE International Conference on Computer Animation and Social Agents (CASA) 2014 and the Co-Editor of the “Data-Driven 3D Facial Animation” book (Springer, 2007).

Stephen Cox trained firstly as a physicist and then as an electronic engineer, and began his career at the UK Government Communications Centre developing signal-processing algorithms. He joined the Speech Applications Division at BT Laboratories UK to work on speech recognition and then worked at the Speech Research Unit at RSRE, Malvern, UK, where he researched into adaptation of speech recognition algorithms to new speakers. He returned to BT to lead a team of researchers developing speech recognition algorithms for use on the UK telephone network. He joined the School of Computing Sciences at UEA as a lecturer and was appointed professor in 2003. He was an invited consultant at AT&T Bell Labs, New Jersey in 1994, a visiting scientist at Nuance Communications Inc., CA, in 2000, and an invited researcher at Apple Inc. in 2010. His research interests include several aspects of computational speech processing (speaker adaptation, confidence measures for speech recognition, automatic call-routing), automatic lip-reading, music-processing and audio event identification. He is the author of over 80 papers in these fields, acts as a consultant to the EU and the UK and Finnish governments. He is a senior member of the IEEE and was Technical Chair of the 2009 Interspeech Conference.