Introduction to the Special Issue on MMAC: Multimodal Affective Computing of Large-Scale Multimedia Data

Sicheng Zhao, *Columbia University, New York, NY, 10027, USA*

Min Xu, *University of Technology Sydney, Sydney, NSW, 2007, Australia*

Qingming Huang, *University of Chinese Academy of Sciences, Beijing, 100864, China*

Björn W. Schuller, *Imperial College London, London, SW7 2BX, U.K., and also with the audEERING, 82205 Gilching, Germany, and also with the University of Augsburg, 86159 Augsburg, Germany*

Humans are emotional creatures. Emotion is present everywhere in our daily life and plays a vitally important role in our decision-making process. There are roughly two categories of modalities that are widely used by humans to express emotions: explicit affective cues and implicit affective stimuli. On the one hand, to distinguish whether an emotion is induced, one direct way is to check the physical changes of involved humans, such as facial expression, speech, action, gait, and physiological signals (e.g., electroencephalogram). These explicit affective cues can be directly observed and collected from an individual with specific sensors. On the other hand, the rapid development of digital photography and social networks has enabled humans to share their lives and expressing their opinions online using implicit affective stimuli, such as text, images, audios, and videos. The user-generated content provides an implicit solution to analyze humans' emotions. Affective computing of explicit and/or implicit large-scale multimedia data is rather challenging due to the following reasons. First, the development of affective analysis is constrained by the affective gap between low-level affective features and high-level emotions. Second, emotion is a subjective concept, and thus affective analysis involves multidisciplinary understanding of human perceptions and behaviors. Finally, emotions are often jointly expressed and perceived through multiple modalities. Multimodal data fusion and complementation need to be explored. This Special Issue (SI) of *IEEE MultiMedia* aims to gather high-quality and cutting-edge contributions reporting the most recent progress on multimodal affective computing of large-scale multimedia data. It targets a mixed audience of researchers and product developers from several communities: multimedia, machine learning, psychology, artificial intelligence, etc.

After rigorous peer review, ten articles were accepted, each of which was reviewed by at least two (usually three) reviewers and went through at least two rounds of reviews. The topics, goals, and contributions of the accepted articles are summarized below.

The initial four articles of this SI focus on **emotion classification, key region localization, and emotion enhancement of single-modal data**. In the article "Facial expression recognition with multi-scale graph convolutional networks," Rao *et al.* propose a novel multiscale graph convolutional network (GCN) for facial expression recognition (FER). To overcome the problems of redundant information and data bias faced in existing convolutional neural networks (CNNs)-based deep FER frameworks, facial landmarks that can describe the location and shape of different facial parts are extracted to efficiently represent facial expressions. Based on the detected facial landmarks, undirected graphs are generated and sent to multiscale GCN to extract graph representations for expression classification. Inspired by the destruction and construction learning and human visual perception, Xia *et al.* propose a simple and effective framework to learn discriminative features in local key facial regions for FER in the article "Destruction and reconstruction learning for facial expression recognition." Besides, the commonly used feature extraction module and the classification module, the proposed ADC-

1070-986X © 2021 IEEE
Digital Object Identifier 10.1109/MMUL.2021.3088548
Date of current version 30 June 2021.

Net also contains a destruction module and a reconstruction module, which are used, respectively, to destroy the global structure of the input image with random shuffle and to restore the global structure of the original input image from the shuffled image. In the article "A magnitude and angle combined optical flow feature for micro-expression spotting," Guo *et al.* study an accurate microexpression (ME) spotting and locating problem. A novel ME dataset SDU2 is constructed with hybrid expressions of 1,602 video clips labeled by professional psychologists. Further, an ME spotting method is proposed based on a magnitude- and angle-combined optical flow feature to exploit the angle information, which can effectively detect the local facial movements. Differently, in the article "Sentiment-aware emoji insertion via sequence tagging," Lin *et al.* focus on a sentiment-aware emoji insertion task to enhance the emotion representation in text. A large-scale emoji insertion corpus, named MultiEmoji, is constructed, including 420,000 English posts with at least one emoji per post. The emoji insertion process is formulated as a sequence tagging task. A BERT-BiLSTM-CRF model is applied to predict multiple emojis and their positions in a sentence.

The subsequent four articles investigate new solutions for **traditional emotion classification from multimodal data**. The article *"Multimodal event-aware network for sentiment analysis in tourism"* by Wang *et al.* proposes an event-aware multitask end-to-end framework to simultaneously predict the travel events and sentiment from multiple modalities, including text and images. Besides, the single-modal content, cross-modal relations are explored by an attention mechanism to learn discriminative multimodal representations. The article "End-to-end learning for multimodal emotion recognition in video with adaptive loss" by Huynh *et al.* studies an emotion recognition problem in video from visual–audio–textual modalities. A lightweight deep architecture is proposed to efficiently extract features for real deployment. The interaction between different modalities is modeled by both an attention strategy and an adaptative loss, which are employed to, respectively, adjust the contribution of different modalities and the network's gradient. In the article "Multimodal and context-aware emotion perception model with multiplicative fusion," Mittal *et al.* investigate context-aware multimodal emotion recognition. Besides human co-occurring modalities (such as facial, audio, textual, and pose/gaits), another two interpretations of context (e.g., background semantic information and sociodynamic interactions among people) are jointly combined by multiplicative fusion, which can focus on the more informative input channels and suppress others for every incoming datapoint. Instead of classifying one multimodal sample into one of the basic emotion categories, in the article "Emotion detection for conversations based on reinforcement learning framework," Huang *et al.* propose to keep track of the gradual emotional changes from every utterance throughout the conversation and uses this information for each utterance's emotion detection. Specifically, a novel reinforcement learning framework is designed to define an agent, the states, and corresponding actions. The progressive emotional interaction process is formulated as a sequential decision problem.

The last two articles are about **emerging affective computing tasks from multimodal data**. Instead of exploring the complementary information by feature fusion, Wu *et al.* explicitly models the incongruity between different modalities for multimodal sarcasm detection in the article "Modeling incongruity between modalities for multimodal sarcasm detection." Concretely, an incongruity-aware attention network (IWAN) is proposed to focus on the word-level incongruity between modalities via a scoring mechanism, which can assign larger weights to words with incongruent modalities. In the article "Implicit emotion relationship mining based on optimal and majority synthesis from multimodal data prediction," Wang *et al.* propose to mine implicit emotion relationship in multimodal data, including emotion distribution, confusion, and transfer. Optimal prediction emotion synthesis and majority prediction emotion synthesis are proposed to synthesize the outputs of multiple emotion classification models. Based on the emotion synthesis results and relative entropy and Jensen–Shannon divergence, implicit emotion relationship is obtained, which is finally applied in topic scene.

In conclusion, the accepted ten articles in this SI cover several important topics from different aspects of multimodal affective computing of large-scale multimedia data. We do believe that these high-quality and cutting-edge research articles will promote the development of multimedia and affective computing fields and bring insightful motivations on future investigations for related researchers and engineers. We would like to sincerely thank Prof. Shu-Ching Chen, the Editor-in-Chief of *IEEE MultiMedia*, and Prof. Mohan Kankanhalli, an Associate Editor-in-Chief, for their kind guidance and support of this SI. The timely and effective help from several IEEE staff, such as Jessica Ingle, Kimberly Sperka, and Sara Scudder, is highly appreciated. We also thank the professional and generous reviewers for their efforts in providing high-quality and suggestive reviews on time. Finally, we thank all the authors who submitted their excellent work on affective computing to this SI.

SICHENG ZHAO (Senior Member, IEEE) is currently a postdoc research scientist at Columbia University, New York, NY, USA. From 2013 to 2014, he was a visiting scholar at the National University of Singapore, Singapore, from 2016 to 2017, a postdoc research fellow at Tsinghua University, Beijing, China, and from 2017 to 2020, a postdoc research fellow at the University of California, Berkeley, Berkeley, CA, USA. His research interests include affective computing, multimedia, and computer vision. Zhao received the PhD degree from the Harbin Institute of Technology, Harbin, China, in 2016. Contact him at schzhao@gmail.com.

MIN XU is currently an associate professor with the School of Electrical and Data Engineering, Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW, Australia. She has authored or coauthored more than 100 research articles in prestigious international journals and conferences. Her research interests include multimedia data analytics and computer vision. Xu received the BE degree from the University of Science and Technology of China, Hefei, China, the MS degree from the National University of Singapore, Singapore, and the PhD degree from the University of Newcastle, Callaghan, NSW, Australia. She is an associate editor of the *Journal of Neurocomputing* (Elsevier). She has been invited to be a member of the program committee for many international conferences. Contact her at Min.Xu@uts.edu.au.

QINGMING HUANG (Fellow, IEEE) is currently a professor at the University of Chinese Academy of Sciences, Beijing, China, and an adjunct research professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 300 academic articles in prestigious international journals and top-level international conferences. His research interests include multimedia computing, image processing, computer vision, and pattern recognition. Huang received the bachelor's degree in computer science and the PhD degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He was the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, and PSIVT. He is also an associate editor of the *IEEE Transactions on Circuits and Systems for Video Technology* and *Acta Automatica Sinica*, and a reviewer of various international journals, including the *IEEE Transactions on Pattern Analysis and Machine*, *IEEE Transactions on Image Processing*, and *IEEE Transactions on Multimedia*. He is a fellow of the IEEE. Contact him at qmhuang@ucas.ac.cn.

BJÖRN W. SCHULLER (Fellow, IEEE) heads the Group on Language Audio & Music (GLAM), Imperial College London, London, U.K., is the CEO of the audio intelligence company audEERING, Gilching, Germany, and a full professor in computer science at the University of Augsburg, Augsburg, Germany. He further holds a visiting professorship at the Harbin Institute of Technology, Harbin, China. Previous positions of his include full professor at the University of Passau, Germany, and visiting professor, associate, and scientist at VGTU, Lithuania, University of Geneva, Switzerland, Joanneum Research, Austria, Marche Polytechnic University, Italy, and CNRS-LIMSI, France. Schuller received thediploma, doctoral, and habilitation degrees from TUM, Munich, Germany, all in electrical engineering and IT. His more than 750 technical publications including more than 100 journal articles (more than 18000 citations, h-index 66) focus on machine intelligence for affective multimedia analysis. He is a Fellow of the IEEE, President-Emeritus of the AAAC, the Editor in Chief of the *IEEE Transactions on Affective Computing*, Associate Editor for the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, and *IEEE Signal Processing Letters* among other associate and multiple guest editorships, and a General Chair of the oncoming IEEE ACII 2019 and the Technical Chair of Interspeech 2019 among various past according and further roles. He was the recipient of a range of awards including being honored as one of 40 extraordinary scientists under the age of 40 by the World Economic Forum in 2015. In 2017, his company secured the 1st place as "Innovator of The Year" of the Digital Marketing Innovation World Cup. His research has garnered more than 10 million USD in extramural funding. His advisory board activities comprise his role as a consultant of global enterprises such as Barclays, GN, Huawei, and Samsung. Contact him at bjoern.schuller@imperial.ac.uk.