Guest Editorial
Introduction to the Special Section on Intelligent Visual Content Analysis and Understanding

VISUAL content analysis and understanding attract tremendous attention because of its potentially wide range of applications including human activity analysis, automated photo face tagging, multicamera tracking, crowded counting, and biometric security. With recent progress in end-to-end differentiable learning, the accuracy of algorithms has been significantly improved and even outperforms humans in some tasks. In addition, multimodality methods, targeting on making full use of various visual data sources, are further investigated. These developments contribute to the innovations of two core modules for a typical intelligent vision system, i.e., image and video description and recognition, which are critical for the success of the visual content analysis and understanding in more complex and challenging open world.

This Special Section provides insights into these explorations, which consists of six articles on visual representations, seven articles on biometric identification, and six action recognition articles.

Depending on different application requirements, the image and video representations focus on different properties, such as saliency detection and preservation, and rotational invariant and cross-modal interaction.

Saliency detection and preservation have been well explored for natural images and videos in the past decades. The classic saliency models typically derive from specific human perception properties such as center-surround mechanisms, contrast cues, and so on. As entering the deep learning era, various powerful deep neural networks further enable an intelligent vision system to cope with more complex scenarios. However, with the emergence of various new types of visual data, greater challenges constantly appear in succession.

"Multi-exposure decomposition-fusion model for high dynamic range image saliency detection," by Wang *et al.*, presents a two-stage framework to estimate the saliency map toward high dynamic range (HDR) image. The first module decomposes a raw HDR image into multiple low dynamic range (LDR) samples under different exposure values, and a salient region proposal network is applied to all LDR samples to generate their candidate saliency maps. The second module derives the overall saliency map via an uncertainty weighting-based fusion toward all of the previous candidates. "RGBT salient object detection: Benchmark and a novel cooperative ranking approach," by Tang *et al.*, introduces a multimodality image dataset and a cooperative ranking-based salient object detection (SOD) method, which focuses on analyzing the RGB and thermal infrared image pairs with the graph optimization model.

For salient object preservation, "Object detection-based video retargeting with spatial–temporal consistency," by Lee *et al.*, adopts object detector and tracker to extract the regions of interest (RoI) at first. Then, all columns covered by the RoIs remain unchanged. Only the surrounding background areas are relocated into the resized frame, where the hole columns are filled via directional interpolation.

Rotational invariant is another highly preferred property for image and video description, which is important in shape and texture features. "Image description with polar harmonic Fourier moments," by Wang *et al.*, first constructs a new type of radial basis functions (RBFs) to suppress the numerical instability. Then, the polar harmonic Fourier moments are computed based on these new RBFs, whose polynomial is invariant with respect to the axis rotations.

Cross-modal interaction is critical for describing the visual data along with the corresponding textual information. "Video dialog via multi-grained convolutional self-attention context multi-modal networks," by Gu *et al.*, employs a multigrained convolutional self-attention context network to learn the joint representations of visual and textual information, which captures both the element and segment-level interactions between the video and its dialog history. "Multimodal transformer with multi-view visual representation for image captioning," by Yu *et al.*, extends the textual transformer model to a multimodal transformer model, which simultaneously captures intra- and intermodal interactions in a unified attention block. Meanwhile, the input image is fed to multiple backbones to extract multiview visual features, which is helpful for preserving fine-grained semantics.

In addition, as the most widespread biometric identification tasks, several face recognition and person re-identification articles are selected to introduce the current progress and trend.

Face recognition has been studied for many decades and become more significant due to its wide range of applications. Various solutions have been proposed to improve the

Digital Object Identifier 10.1109/TCSVT.2020.3031416

performance of face recognition, more specifically in terms of accuracy and efficiency. Toward high computing efficiency in video-based face recognition, "Face recognition based on videos by using convex hulls," by Cevikalp *et al.*, uses kernelized support vector data description to extract representative samples (support vectors in kernel space) that formed the sample set boundaries, which approximates the original sample set and significantly reduces the redundancy. Then, they propose a binary hierarchical decision tree approach to improve the speed up to 20 times. In addition, a new video-based face recognition dataset is introduced for large-scale set-based face recognition.

To improve the accuracy when facing large variations in the biometric data characteristics, "A double-deep spatio-angular learning framework for light field-based face recognition," by Sepas-Moghaddam *et al.*, proposes a learning framework to embed spatio-angular information provided by Light Field camera to assist face recognition. In detail, a pre-trained VGG-16 convolutional neural network is used to obtain VGG-face descriptions for each subaperture (SA) image corresponding to each observation angle. And a long short-term memory (LSTM) network is used to integrate VGG-face descriptions in the angular domain. Various SA image selection strategies are further investigated to obtain minimal SA image sets that contain sufficient angular information to provide accurate results.

"Modeling and optimizing of the multi-layer nearest neighbor network for face image super resolution," by Chen *et al.*, tries to improve face recognition accuracy by super-resolving the human face image. To this end, a Nearest Neighbor Network is built in the training stage to model the neighbor relationship through similarity between Low-Resolution (LR) human face image patches and High-Resolution (HR) counterparts, which enforces corresponding HR and LR patches close to each other in the neighbor graph. And in the testing stage for an unseen LR image, HR patches are obtained using the first- and second-order neighbors of the corresponding patches in the LR image.

Person Re-Identification refers to matching persons captured by multiply cameras. Since the appearance, illumination and the data modality usually vary from different cameras, the key problem relies on learning the metrics and representations which are invariant to view changes of the same identity and discriminative to different identities. In the deep learning era, metrics and representation learning are achieved by network architecture and loss function design.

"Deeply associative two-stage representations learning based on labels interval extension loss and group loss for person re-identification," by Huang *et al.*, uses the global reinitialization stage and fully-perceptual classification stage to construct a two-stage architecture, where the first stage adopts ResNet50 as a backbone network and is further enhanced with the proposed a 3-D transpose technique. The two-stage network is then optimized with a joint loss function that combining the proposed Labels Interval Extension loss and Group loss. More specifically, the Labels Interval Extension loss uses softmax function to enforce the predicted identities distribution (probabilities to be each identity) close to ground truth label while the group loss is designed to enlarge the distance among different identities. "Three-dimension transmissible attention network for person re-identification," by Huang *et al.*, introduces 3-D transmissible attention mechanism into the network design to transfer channel–spatial attention information between convolution layers and retained holistic attention information. In addition, L2-norm multilabel loss is used to decrease intragroup distance and increase intergroup distance.

In addition to network design, "Top-push constrained modality-adaptive dictionary learning for cross-modality person re-identification," by Zhang *et al.*, discusses the cross-modality person re-identification problem, in which RGB, infrared, and depth cameras are used to capture the human appearance. Facing extremely different appearances under different modality, heterogeneous features from dissimilar modality are projected into a canonical space where modality-specific bias is mitigated by enforcing different modalities of the same person close to each other. In addition, a top-push ranking graph regularization is used in the neighbor graph. "Large margin nearest neighbors classification with privileged information for biometric applications," by He *et al.*, introduces privileged information into the person re-identification problem, where additional depth images associated with the RGB are available only in the training stage. The Subgradient Descent algorithm is adopted to solve the optimization problem with constraints conditions.

Action recognition is another popular field of visual content analysis and understanding, which is challenging due to the complex spatio–temporal variations of different behaviors in diverse video sequences. In the deep learning era, learning powerful and generic spatio–temporal representations is achieved by designing specific network architectures or loss functions.

Toward accurate and efficient sport video analysis, "Learning to score figure skating sport videos," by Xu *et al.*, proposes a Large scale figure skating dataset and a semantic representation learning framework to learn effective representation for intelligent analysis. In detail, a pretrained network is used to encode the spatial information of each frame, forming a temporal feature sequence. Based on the feature sequence, the temporal semantic information is further encoded by a self-attentive LSTM and Multi-Scale Convolutional LSTM, with the former one responsible for encoding global temporal semantic information and the later one responsible for encoding local temporal semantic information.

To explore the complementary properties between the hand-crafted shallow feature representation and deep features, "Discriminative multi-view subspace feature learning for action recognition" by Sheng *et al.*, proposes a subspace learning model for effective shallow and deep feature fusion. To be specific, Fisher Vector is used to encode improved dense trajectories for shallow representation and two-stream model is used to generate deep features. The shallow and deep features are then projected into a shared subspace for effective fusion, and the projection parameters are learned with the guidance of a reconstruction error and classification error.

To improve the performance of 3-D human pose estimation, especially under the context of diverse viewpoints, the paper “View invariant 3D human pose estimation,” by Guo *et al.*, proposes a marvelous view-invariant hierarchical correction network (VI-HC) along with a view-invariant discriminative network (VID) to enhance the representations of all joints on body. Specifically, VI-HC transforms initial 3D pose inputs to consistent views and further refines them at the global body, local body parts level, respectively. VID applies adversarial learning to differentiate between estimated 3-D poses and real 3-D poses to avoid implausible results. Unlike previous methods, this paper takes insight into the problem of multiple viewpoints. And the designed VI-HC and VID can help learn more robust representations under such situation.

Loss function is one of the indispensable components in convolutional neural networks, and different loss functions can assist CNN to learn different feature representations. The paper, “A novel CNN training framework: Loss transferring,” by Liang *et al.*, introduces a novel training framework by transferring the knowledge learned via one loss function to another. This loss transferring strategy can incorporate different loss functions to improve the generalization ability of the visual representation learned by CNN. They first propose two basic guides to construct an appropriate loss function set. Following the basic guides, a near classifier hyper-plane loss is proposed to reduce the intraclass variance. Then, the authors design two training strategies to improve the training efficiency of CNN model. Thorough experimental results on four benchmark datasets demonstrate the effectiveness of this model.

The last two articles focus on designing robust recognition algorithms. Different kinds of tasks may require focusing on different aspects of the video. Therefore, a simple representation learning method is not capable of dealing with various video analysis tasks. To achieve satisfying performance, effective recognition methods with specifically designed learning strategies are urgently needed for diverse tasks.

Early action recognition has become a popular topic in research of recognition, which aims to predict the action label from a partially observed video. The paper, “Early action recognition with category exclusion using policy-based reinforcement learning,” by Weng *et al*., develops an approach to employ diverse information of different negative categories. They introduce a category exclusion mechanism, which is modeled as a mask operation on the classification probability output of a pre-trained early action recognition classifier. Specifically, the authors adapt the policy-based reinforcement learning to train an agent, which is used to generate a series of binary masks to exclude interfering negative categories during action execution. Cooperated with a sequential classifier, the full model could utilize complementary information of time dynamic in an action sequence, achieving much higher recognition accuracy than previous works.

The article “Attention-driven loss for anomaly detection in video surveillance,” by Zhou *et al.*, describes a novel video anomaly detection method by introducing an attention mechanism to resolve the imbalance between foreground objects and stationary background in a video. The model is composed of an encoder to extract features and a decoder to predict the next frames. A single attention map is first computed to capture the temporal motion cues across the video frames while suppressing the stationary background pixels. Then, an attention-driven content loss is proposed to alleviate the data bias issue and enable the network to focus on the regions of interest. Furthermore, an attention driven gradient difference loss is introduced to sharpen the image prediction. The experimental results on Avenue, Ped1, and Ped2 datasets validate the effectiveness of the model compared to state-of-the-art methods.

Recent years have witnessed the popularization of numerous intelligent vision systems. Although significant progress has been made in the past decades, there are still great deals of challenges in this field due to various emerging visual media and our limited knowledge toward human visual recognition. Limited by the space, this Special Section only catches a few explorations from the leading researchers in visual content analysis and understanding. Hopefully, these works could throw light on your research to understand the pleasant and enjoyable visual world.

HONGLIANG LI, *Lead Guest Editor*
School of Information and
Communication Engineering
University of Electronic Science and
Technology of China
Chengdu 611731, China
e-mail: hlli@uestc.edu.cn

LU FANG, *Guest Editor*
Department of Electronic Engineering
Tsinghua University
Beijing 100084, China
e-mail: fanglu@sz.tsinghua.edu.cn.

TIANZHU ZHANG, *Guest Editor*
Department of Automation
University of Science and Technology
of China
Hefei 230026, China
e-mail: tzzhang@ustc.edu.cn

Hongliang Li (Senior Member, IEEE) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005.

From 2005 to 2006, he joined the Visual Signal Processing and Communication Laboratory (VSPC), Chinese University of Hong Kong (CUHK), as a Research Associate, where he was a Post-Doctoral Fellow, from 2006 to 2008. He is currently a Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. He has authored or coauthored numerous technical articles in well-known international journals and conferences. He is a Co-Editor of *Video Segmentation and its Applications*. His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia processing. He is involved in many professional activities. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology and *Journal of Visual Communications and Image Representation* and the Area Editor of *Signal Processing: Image Communication*. He served as a Technical Program Chair for VCIP2016 and PCM2017, the General Chair of the ISPACS 2010, the Publicity Chair of IEEE VCIP 2013, the Local Chair of the IEEE ICME 2014, and TPC Member in a number of international conferences, e.g., ICME 2013, ICME 2012, ISCAS 2013, PCM 2007, PCM 2009, and VCIP 2010.

Lu Fang (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China in 2007 and the Ph.D. degree from the Hong Kong University of Science and Technology in 2011. She is currently an Associate Professor with Tsinghua University. Her research interests include computational photography and 3D vision. She has received the Multimedia Rising Star Award in ICME 2019 and the Best Student Paper Award in ICME 2017. She is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).

Tianzhu Zhang (Member, IEEE) received the B.E. degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

From 2011 to 2013, he was a Research Fellow/Scientist with the Advanced Digital Sciences Center (ADSC), a joint research center between the University of Illinois at Urbana–Champaign (UIUC) and the Agency for Science, Technology, and Research (A∗STAR), Singapore. From 2013 to 2018, he was an Assistant/Associate Professor with the Institute of Automation, Chinese Academy of Sciences. Since 2019, he has been a Professor with the School of Information Science and Technology, University of Science and Technology of China. He has authored or coauthored more than 100 journals and conference papers in these areas, including more than 40 IEEE/ACM Transactions papers (TPAMI/TIP/IJCV) and more than 30 top-tier computer vision conference papers (ICCV/CVPR/ECCV). His research interests include computer vision, pattern recognition, multimedia computing, and machine learning. He was a recipient of the Best Paper awards at the ACM MM 2016 and ChinaMM 2017, and one top 10% paper award at the ICIP 2013. He received the Outstanding Reviewer Award in MMSJ 2013, ECCV 2016, and CVPR 2018. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT) and *Neurocomputing*, and the Publicity Chair of ICIMCS 2015 and ACM Multimedia Asia 2019. He served/serves as the Area Chair for ECCV 2020, ACM Multimedia 2020, CVPR 2020, ICCV 2019, ACM Multimedia 2019, WACV 2018, ICPR 2018, and MVA 2017.