Multimed Tools Appl (2018) 77:29209–29211
https://doi.org/10.1007/s11042-018-6341-6

GUEST EDITORIAL

Guest Editorial: Spatio-temporal Feature Learning for Unconstrained Video Analysis

Yahong Han[1] · Liqiang Nie[2] · Fei Wu[3]

Published online: 10 July 2018

With the development of mobile Internet and personal devices, we are witnessing an explosive growth of video data on the Web. This has encouraged the research of video analysis, which can boost the development of techniques for the management and applications of videos, such as video retrieval, video classification, action recognition, event detection, and video captioning etc. Compared to the trimmed videos in the open benchmark datasets, most of the real-world videos are unconstrained video data. Firstly, as captured under different conditions, unconstrained videos usually have large intra-class differences. Secondly, as captured by different devices and people, unconstrained videos may own more variants in quality. How to develop a robust feature or representation is the key problem in unconstrained video analysis. Previous studies mainly focus on the hand-crafted video descriptors, e.g., STIP, MoSIFT, Dense Trajactory etc. The success of these descriptors lies in the simultaneously incorporating spatial description of each frame and temporal consistency of successive frames. Recently, researchers have tried to learn video representations from deep ConvNets, where the promising progresses were obtained owing to the breakthrough in the appropriately pooling or encoding of temporal information of video sequences in the deep neural networks, e.g., Two-Stream and TDD. As the visual content and temporal consistency of unconstrained videos are more complex, there are still challenges in video analysis and practical applications.

Spatio-temporal feature learning plays an important role in various applications of video analysis, e.g., video classification, video summarization, visual attention prediction, and video

✉ Yahong Han
yahong@tju.edu.cn

Liqiang Nie
nieliqiang@gmail.com

Fei Wu
wufei@cs.zju.edu.cn

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China

[2] School of Computer Science and Technology, Shandong University, Jinan, China

[3] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

question answering. In this issue, four papers investigate how to learn an effective video representation by the combination of spatial and temporal cues. The paper 'Tensor Learning and Automated Rank Selection for Regression-based Video Classification' proposes a tensor-based logistic regression learning algorithm to effectively exploit underlying space-time structural information in video sequences, in which the weight parameters are regarded to be a tensor, calculated after the CP tensor decomposition. Though deep learning methods have dominated the area of video feature learning, tensor learning is valuable to be explored, as tensor is a natural structure to represent the space-time structural information in video sequences. The paper 'Hybrid Convolutional Neural Networks and Optical Flow for Video Visual Attention Prediction' developed a deep-learning framework to learn a hybrid spatial temporal video feature from CNN and optical flow. Their method shows good performance in applications of visual attention prediction. The paper 'Foveated Convolutional Neural Networks for Video Summarization' tries a novel idea to integrate gaze information into a deep learning framework. Foveated images are constructed based on subjects' eye movements to represent the spatial information of the input video. Multi-frame motion vectors are stacked across several adjacent frames to convey the motion clues. Experimental results demonstrate such a new scheme of video feature learning can lead to good performance of video summarization. Visual Question Answering (VQA) is a recent hot topic and the challenge lies in that, in most cases, it requires reasoning over the connecting between visual content and languages. The paper 'Remember and Forget Video and Text Fusion for Video Question Answering' explore the video feature learning for video question answering. They developed a new memory network to well fuse spatio-temporal video features with textual information.

As this issue is also dedicated to serve as a forum of practical applications for unconstrained video analysis, we also solicited five papers towards different practical applications, especially under the unconstrained settings. The paper 'Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features' explores an important real-world applications of smoke detection in videos, which may be applied to reduce the fire losses. Besides a new framework for spatio-temporal feature learning, they developed an effective detection method with real-time efficiency. The paper 'Video logo removal detection based on sparse representation' targets the forged video content, which has potential applications in legal authorities of videos. They presented a video forensics framework for logo removal detection, which mainly contains two stages: the removal traces detection and the removal region location. Experimental results show promising results on their video logo removal dataset. Different from the logo detection in videos, the paper 'Spatiotemporal Text Localization for Videos' focuses on spatiotemporal text localization in videos. Their method exploits the temporal redundancy of text to increase the detection efficiency for videos. The other two papers targets fancy applications of videos, i.e., production of movie trailers and virtual training system for ancient chinese architecture. The paper 'Embedded learning for computerized production of movie trailers' use CNN to extract features of candidate frames from the film by a rank-tracing technique. Then a SURF algorithm is utilized to match the frames of the movie with the corresponding trailer. In the paper 'Design and Development of a Maintenance and Virtual Training System for Ancient Chinese Architecture', a building information model (BIM) and virtual reality (VR) and video analyzing technology are combined to develop a maintenance and virtual training system for ancient architecture.

In this issue, we also include a paper about video benchmark, i.e., 'MMA: A Multi-view and Multi-modality Benchmark Dataset for Human Action Recognition'. The proposed dataset MMA consists of 7080 action samples from 25 action categories, including 15 single-subject actions and 10 double-subject interactive actions in three views of two different scenarios. Authors systematically benchmark the state-of-the-art approaches on MMA by different temporal-spatial feature representations. Moreover, MMA follow the design principle towards unconstrained settings of significant intra-class variations, occlusion issues, views and scene variations, and multiple similar action categories.

Of the 25 papers submitted to this issue, 10 were eventually accepted after a stringent peer review process. These 10 papers cover a wide range of methods and applications about spatio-temporal feature learning for unconstrained video analysis. As videos in practical applications, e.g. surveillance videos, may contain more noises and variants. It is more difficult to train a general model to capture the spatial-temporal cues in different practical applications. Therefore, this special issue encourages practical methods for real-world unconstrained or untrimmed video data. We hope this issue appeal to both the experts in the field as well as to those who wish a snapshot of the current breadth of practical video analysis.

Acknowledgments This work is supported by the NSFC (under Grant U1509206,61472276) and Tianjin Natural Science Foundation (no. 15JCYBJC15400).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.