Multimedia Systems (2022) 28:361–362
https://doi.org/10.1007/s00530-021-00855-w

EDITORIAL

Editorial

Meng Liu[1] · Yan Yan[2] · Tian Gan[3] · Hua Huang[4] · Mohan Kankanhalli[5]

Published online: 30 October 2021

We are living in the era of multimedia, a tremendous number of videos, images, and texts are generated, published, and spread daily. In other words, multimedia data have been becoming an indispensable part of today's big data. In fact, the large-scale multimedia data have raised challenges and opportunities for developing intelligent multimedia systems, such as retrieval, recommendation, recognition, categorization, and generation systems. Although shallow learning has achieved some progress, its processing capacity on large-scale data is still limited. Meanwhile, deep-learning algorithms have enabled the development of highly accurate systems and have become a standard choice for analyzing different types of data. For instance, convolutional neural networks have demonstrated high capability in image classification, recurrent neural networks are widely exploited in modeling temporal sequence in NLP. Inspired by this, we are keen on applying deep-learning techniques to boost the performance of multimedia analysis tasks. This special issue seeks recent advances in the deep learning-based multimedia analytics and relatively new areas.

Submissions came from an open call for paper and with the assistance of professional referees. Ten papers were finally selected after at least two rounds of rigorous peer review. These accepted papers cover several popular topics of multimedia analysis, including point cloud analysis, visual question answering, 3D human pose estimation, video keyframe extraction, image forgery detection, and cross-media retrieval.

The paper entitled "Hybrid features and semantic reinforcement network for image forgery detection", co-authored by Haipeng Chen, Chaoqun Chang, Zenan Shi, and Yingda Lyu, proposes a hybrid features and semantic reinforcement network to detect and localize image manipulation regions at the pixel level, including copy-move, splicing, and object removal. It first hybrids consolidated features from rotating residual units and resampled features extracted from the LSTM, and then, designs the semantic reinforcement between encoding and decoding network to improve usage of shallow layers semantic information. This scheme is shown to be better than several state-of-the-art methods on NIST16, COVERAGE, and CASIA.

The paper entitled "Pinyin-to-Chinese conversion on sentence-level for domain-specific applications using self-attention model", co-authored by Shufeng Xiong, Li Ma, Ming Cheng, and Bingkun Wang, proposes a sequence to sequence learning approach for the Pinyin-to-Chinese task without pinyin segmentation and candidate generation. It leverages the self-attention mechanism on the encoder to map an input to intermediate features, and then, adopts the softmax layer to predict the Chinese character at each position corresponding to each input pinyin character. The experimental results demonstrate the advantage of the proposed scheme.

The paper entitled "Key frame extraction based on global motion statistics for team-sport videos", co-authored by Yuan Yuan, Zhe Lu, Zhou Yang, Meng Jian, Lifang Wu, Zeyu Li, and Xu Liu, proposes a global motion statistics-based scheme to extract key frames from team-sport videos, by considering global motion information. This paper builds a dataset, SportKF, for team-sport videos key frame extraction. Experimental results show that this method offers better performance than several state-of-the-art ones.

✉ Meng Liu
mengliu.sdu@gmail.com

Yan Yan
tom.yan.555@gmail.com

Tian Gan
gantian@sdu.edu.cn

Hua Huang
11112020041@bnu.edu.cn

Mohan Kankanhalli
mohan@comp.nus.edu.sg

[1] Shandong Jianzhu University, Jinan, China

[2] Texas State University, San Marcos, USA

[3] Shandong University, Jinan, China

[4] Beijing Normal University, Beijing, China

[5] National University of Singapore, Singapore, Singapore

The paper entitled "3D human pose estimation with multi-scale graph convolution and hierarchical body pooling", co-authored by Ke Huang, Tianqi Sui, and Hong Wu, proposes a GCN-based network for single-frame 3D human pose estimation, which utilizes multi-scale graph convolution module to aggregate features of neighbors at different distances and presents hierarchical-body-pooling to extract and share body-level and body-part-level context information. The proposed approach is shown to be better than several state-of-the-art but with much fewer parameters.

The paper entitled "Skip-attention encoder-decoder framework for human motion prediction", co-authored by Ruipeng Zhang, Xiangbo Shu, Rui Yan, Jiachao Zhang, and Yan Song, proposes the skip-attention encoder–decoder framework to model human motion dependences in spatiotemporal space, by utilizing the encoder and decoder to encode the observed motions and decode the predicted motions, respectively. This scheme is shown to be better than several state-of-the-art methods in both quantitative and qualitative results.

The paper entitled "DBFC-Net: a uniform framework for fine-grained cross-media retrieval", co-authored by Qiong Wang, Youdong Guo, and Yazhou Yao, proposes a double branch fine-grained cross-media network, which uses the media-specific information to construct the common features by a uniform framework. Moreover, it devises a distance metric (cosine +) for fine-grained cross-media retrieval. Experiments on publicly available datasets demonstrate the effectiveness of this proposed approach.

The paper entitled "Structure injected weight normalization for training deep networks", co-authored by Xu Yuan, Xiangjun Shen, Sumet Metha, Teng Li, Shiming Ge, and Zhengjun Zha, proposes deep structural weight normalization methods to inject the network structure measurements into the weight normalization to fully acknowledge the data propagation through the neural network. The proposed method is shown to be able to reduce the number of trainable parameters while guaranteeing high accuracy, whereas accelerate the convergence while improving the performance of deep networks.

The paper entitled "Question-relationship guided graph attention network for visual question answer", co-authored by Rui Liu, Liansheng Zhuang, Zhou Yu, Zhihao Jiang, and Tian Bai, proposes a question-relationship guided graph attention network for VQA, which utilizes three graph encoders with diverse relationships to capture high-level features of images. This scheme is shown to be better than other interpretable visual context structures.

The paper entitled "Direction-induced convolution for point cloud analysis", co-authored by Yuan Fang, Chunyan Xu, Chuanwei Zhou, Zhen Cui, and Chunlong Hu, proposes a direction-induced convolution to obtain the hierarchical representations of point clouds and then boost the performance of point cloud analysis. Experiments on three benchmark datasets, ModelNet40, ShapeNet Part, and S3DIS, demonstrate the effectiveness of the proposed method on both point cloud classification and semantic segmentation tasks.

The paper entitled "Assessing learning engagement based on facial expression recognition in MOOC's scenario", co-authored by Junge Shen, Haopeng Yang, Jiawei Li, and Zhiyong Cheng, presents a framework to assess the learning engagement of online learners in the MOOC scenario via facial expression recognition. Moreover, to recognize the facial expression, a domain adaptation network is advanced, which is suitable for the MOOC scenario. The experimental results on JAFFE, ck+, and RAF-DB demonstrate the effectiveness of the proposed method.

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