International Journal of Computer Vision (2021) 129:3171–3173
https://doi.org/10.1007/s11263-021-01530-3

EDITORIAL

Guest Editorial: Special Issue on Deep Learning for Video Analysis and Compression

Dong Xu[1] · Rama Chellappa[2] · Luc Van Gool[3,4] · Guo Lu[5]

Accepted: 16 September 2021 / Published online: 15 October 2021

1 Introduction

Due to the rapid popularization of digital cameras and mobile phone cameras, there is an increasing research interest in developing next-generation technologies for storing, transmitting, indexing, and understanding various types of videos including movies, surveillance videos, web videos and personal videos. Deep learning technologies have demonstrated excellent performance in a broad range of video content analysis tasks such as activity recognition and video event recognition, video-based biometrics, and video captioning as well as video question and answering. Meanwhile, deep video compression has become a new research direction in visual data compression, and the recent deep video compression technologies have achieved promising results on the benchmark datasets. The goal of this special is to capture the latest research progress in all areas related to deep learning for video analysis and compression.

Among twenty-three submissions, ten deep learning articles covering a broad range of topics on video analysis and compression were accepted. These articles can be grouped into three categories: (1) deep learning for video analysis; (2) deep learning for visual data processing and compression; and (3) joint optimization of image/video analysis and compression. Below, we briefly summarize the articles selected for publication in this special issue.

2 Organization and Overview

2.1 Deep Learning for Video Analysis

The article entitled "A Coarse-to-Fine Framework for Resource Efficient Video Recognition" introduces a new video recognition method referred to as LiteEval by adopting the coarse-to-fine strategy. LiteEval consists of a coarse recurrent neural network (RNN), a fine RNN and a conditional gating module, in which the low-cost features (i.e., the coarse features) are used together with historical information to automatically decide whether computationally expensive features (i.e., the fine features) are still required for processing the current frame. Extensive experiments on three video benchmarks demonstrate LiteEval achieves promising recognition results while using much less computation resources under both online and offline settings.

The article "Context and Structure Mining Network for Video Object Detection" presents a new network to better aggregate features for more accurate video object detection. After encoding spatial–temporal context information into the object features, Han et al. additionally propose a divide-and-match strategy to exploit the structure information of objects, which can well cope with the object pose misalignment and occlusion issues. When compared with the state-of-the-art algorithms, the proposed approach achieves better video object detection results on the ImageNet VID dataset.

In the article "SODA: Weakly Supervised Temporal Action Localization Based on Astute Background Response and Self-Distillation Learning", Zhao et al. propose two new strategies for weakly supervised temporal action localization. The astute background response strategy aims to alleviate the over-localization issue by distinguishing the disturbing background frames from those containing true action instances,

✉ Dong Xu
dong.xu@sydney.edu.au

[1] School of Electrical and Information Engineering, The University of Sydney, Building J03, Sydney, NSW 2006, Australia

[2] Departments of Electrical and Computer Engineering and Biomedical Engineering, Johns Hopkins University (JHU), 3400 North Charles Street, Baltimore, MD 21218, USA

[3] Department of Information Technology and Electrical Engineering, ETF C 117, Sternwartstrasse 7, 8092 Zurich, Switzerland

[4] ESAT Department, KU Leuven, Kardinaal Mercierlaan 1, 3001 Leuven, Belgium

[5] School of Compute Science and Technology, Beijing Institute of Technology, Room 305, Software Building, Beijing 100081, China

while the self-distillation learning strategy aims at addressing the under-localization issue by learning one master network and multiple auxiliary networks. Extensive experiments on three benchmark datasets demonstrate the effectiveness of the newly proposed approach SODA for weakly supervised temporal action localization.

In the article "Deep Trajectory Post-Processing and Position Projection for Single & Multiple Camera Multiple Object Tracking", Ma et al. propose a Siamese Bi-directional Gated Recurrent Units (GRU) structure for object tracking, in which two networks called Cleaving Network and Re-connection Network are developed for trajectory post-processing. For multiple camera tracking, they additionally propose the Position Projection Network to provide accurate temporal-spatial information for trajectory association. The extensive experiments on two benchmark datasets demonstrate the newly proposed approach outperforms the start-of-the-art trackers.

The article "Learning Adaptive Attribute-Driven Representation for Real-Time RGB-T Tracking" presents a new attribute-driven representation network (ADRNet) for real-time RGB-T tracking. By decomposing the tracking challenges into four attributes, Zhang et al. build an effective residual representation for object modeling by using an attribute-driven residual branch for each attribute. They additionally propose an attribute ensemble network to aggregate the representations related to different attributes. Comprehensive experiments on three RGB-T tracking benchmarks demonstrate the newly proposed approach achieves better object tracking performance when compared with the state-of-the-art methods.

The article by Huang et al., entitled "A Decomposable Winograd Method for N-D Convolution Acceleration in Video Analysis", proposes a new Decomposable Winograd method for N-D convolution acceleration, which overcomes the limitations of the conventional Winograd's minimal filtering algorithm when handling large kernel sizes and can thus be readily used for various video analysis applications. Before using the Winograd approach, they decompose the kernels with large size or large stride to a set of small kernels with stride 1, which can simultaneously reduce the number of multiplication operations and keep the numerical accuracy. The experiments demonstrate the newly proposed method supports all types of N-D convolution operations with speedup up to 3.38 times while without degrading the accuracies.

2.2 Deep Learning for Visual Data Processing and Compression

The article "Deep Maximum a Posterior Estimator for Video Denoising" presents a Maximum a Posteriori (MAP)-based video denoising method MAP-VDNet, based on adaptive temporal fusion and a deep image prior. Sun et al. first propose an iterative MAP-based video denoising algorithm and further extend it as a multi-stage implementation, in which both image denoising and image fusion modules can be end-to-end optimized. To address the misalignment issue, a multi-frame fusion strategy is additionally proposed to predict the fusion weights. Comprehensive experiments demonstrate promising denoising performance and excellent generalization capability of the proposed approach.

In the article entitled "Deep Learning Geometry Compression Artifacts Removal for Video-based Point Cloud Compression", Jia et al. propose a new learning framework to improve the point cloud data compression performance by reducing the geometry compression artifacts. Based on the near and far depth fields decomposed from geometry videos, they propose a two-stage method, where a new learning-based pseudo-motion compensation module is first used before exploiting the correlations between near and far depth fields in the second stage. The proposed algorithm can be readily embedded in the V-PCC reference software. Comprehensive experiments demonstrate the effectiveness of the newly proposed approach for geometry artifacts removal.

2.3 Joint Optimization of Image/Video Analysis and Compression

In the article "Just Recognizable Distortion for Machine Vision Oriented Image and Video Coding", Zhang et al. propose a new concept called Just Recognizable Distortion (JRD), which is defined as the maximum distortion after image/video compression that will not reduce the machine vision models' performance to an unacceptable level. They also build a large dataset to facilitate the subsequent research along this direction and propose an ensemble-learning based framework to predict the JRD for various image-based visual recognition tasks under different conditions. Comprehensive experiments demonstrate the effectiveness of the proposed approach for machine vision-oriented image and video coding.

The article entitled "Semantics-to-Signal Scalable Image Compression With Learned Revertible Representations" introduces a new scalable image compression method, in which the partial bitstream and the whole bitstream are decoded for machine vision and human vision tasks, respectively. Based on the lifting structure, Liu et al. first propose a trainable and revertible transform and then each image can be converted into a pyramid of multiple subbands. By jointly optimizing multiple factors including compression ratio, semantic analysis accuracy and signal reconstruction quality, they additionally propose an end-to-end optimized encoding/decoding network for compressing these subbands. The comprehensive experiments demonstrate the newly proposed method achieves semantics-to-signal scalable image compression.

3 Conclusion

The articles selected for publication in this special issue encompass a wide range of research topics related to deep learning for video analysis and compression. We hope that this collection of articles will be helpful for both experts in the related areas and those who want a snapshot of the current breadth of deep video analysis and compression technologies.

Acknowledgements We would like to thank all authors for summitting their outstanding works to this special issue. We are also grateful to the reviewers for their invaluable evaluations and insightful suggestions. We are particularly grateful to the Editor-in-Chiefs Yasuyuki Matsushita, Jiri Matas and Svetlana Lazebnik as well as the editorial staff of IJCV for their support and help throughout the entire process.