SIP (2020), vol. 9, e1, page 1 of 8 © The Authors, 2020.

doi:10.1017/atsip.2019.26

INDUSTRIAL TECHNOLOGY ADVANCES

Toward human-centric deep video understanding

WENJUN ZENG

People are the very heart of our daily work and life. As we strive to leverage artificial intelligence to empower every person on the planet to achieve more, we need to understand people far better than we can today. Human–computer interaction plays a significant role in human-machine hybrid intelligence, and human understanding becomes a critical step in addressing the tremendous challenges of video understanding. In this paper, we share our views on why and how to use a human centric approach to address the challenging video understanding problems. We discuss human-centric vision tasks and their status, highlighting the challenges and how our understanding of human brain functions can be leveraged to effectively address some of the challenges. We show that semantic models, view-invariant models, and spatial-temporal visual attention mechanisms are important building blocks. We also discuss the future perspectives of video understanding.

Keywords: Human-centric, Video understanding, Deep learning

Received 02 March 2019; Revised 11 December 2019

I. INTRODUCTION

Artificial intelligence (AI) is the buzz word in the technology world today. In the past few years, the machine has beaten humans in many ways – facial recognition, image recognition, IQ test, gaming, conversational speech recognition, reading comprehension, language translation, just to name a few.

All these breakthroughs are attributed to three pillars of technological innovations. The first is the availability of the big data, e.g. thousands of hours of annotated speech, and tens of millions of labeled images. The second foundation is the availability of huge computing resources, such as GPU cards and cloud server clusters. On top of these two, we have witnessed the significant progress in advanced machine learning, such as deep learning and reinforcement learning. We are indeed in a golden age of AI.

A) Deep learning has changed the landscape of image understanding

Research estimates that 80–85% of our perception, learning, cognition, and activities are mediated through vision [1]. This signifies the importance of the role that visual intelligence plays in AI.

Deep learning is the engine for AI, and it has changed the landscape of image understanding. Figure 1 shows the error rate performance of the winners of the well-known ImageNet classification competitions over the years. Since 2012, convolutional neural network (CNN)-based approaches have been widely adopted, and the winning systems used increasingly deeper networks. In 2015, Microsoft Research Asia announced a system that beat the human performance and won the competition with a 152-layer network called deep residual network which makes it easier to train deeper networks with shortcut connections between layers [2]. What an amazing progress in a short 4 years, from a system that was far from practical, to a system that beat human performance. This demonstrates the power of deep learning.

Similarly, significant improvements have been achieved using deep learning for object detection and semantic segmentation. We are getting a lot closer to landing computer vision technologies to practice.

B) Video understanding is challenging

Typically, if there is a significant technological development, the image goes first, the video will follow. In fact, there is a huge market for intelligent video analytics in both the enterprise and consumer domains. In addition to traditional public surveillance market, there have been many emerging applications, e.g. in business intelligence, home security, autonomous driving, and storytelling.

In general, videos are much harder to analyze than images. The variety of contents in the video is exponentially larger, making fine-grained vision tasks extremely difficult. There is a huge demand for storage, computing power, and

Microsoft Research Asia, No. 5 Danling Street, Haidian District, Beijing, China

Corresponding author:
W. Zeng
Email: wezeng@microsoft.com

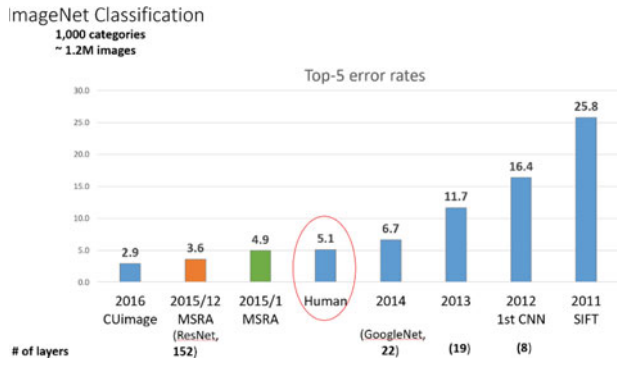

Fig. 1. Performance of the winners of the ImageNet classification competitions over the years.

bandwidth. In some scenarios, real-time processing is a requirement. The labeling for video is particularly costly. In some scenarios, there is a lack of training data. For example, it is more difficult to make surveillance video data available for research, partly due to the privacy issue. In some other cases, positive samples are scarce. These are all the challenges that video understanding faces, making it very difficult to land the video analytics technologies to practice.

Given the challenges in making video understanding technologies practical, we believe it is important to take a human-centric approach to focus on the features that are most critical for bringing the technologies to the market.

II. HUMAN-CENTRIC: WHY AND HOW?

People are the very heart of our daily life and work. In order to serve people better, we need to better understand people, their surroundings, and their relationships: who they are, what they are wearing, what they are doing and saying, how they are feeling, what their intentions are, who they are talking to, among others. Just as people are constantly trying to understand themselves and other people, machines also need powerful tools to help them understand people through multi-modality sensory data in today's societies that are becoming more and more intelligent.

Not surprisingly, people are the main subjects in most videos, and human–computer interaction plays a significant role in human-machine hybrid intelligence that is likely to dominate in practice in the foreseeable future. Therefore, human understanding becomes a critical step in video understanding. In fact, one of the earliest successful applications of computer vision is about human, i.e. the widespread deployment of face recognition. By extension, it is likely that the next breakthrough could come from general human understanding technologies. Therefore, it makes a lot of sense to take a human-centric approach for video understanding. By "human-centric", we mean both focusing on the tasks of understanding humans in video and leveraging what we have understood about how human brain works in the algorithm designs, although what we can do for the latter is still very limited.

After all, the initial goal of AI was to mimic how the human brain works. We need to understand human in general from multiple perspectives, including biological science, neural science, cognitive science, behavior science, and social science, etc. For video understanding, deep learning is a powerful tool that has been shown to have great promise. It is important to understand how the human brain works and then leverage that in the design of deep learning systems. In fact, the human brain's attention mechanism has been successfully applied to neural network designs [3]. Human reasons based on the knowledge they acquire. We are seeing more and more efforts in integrating a knowledge-driven approach with a data-driven approach in deep learning system design [4–6], as will be elaborated on in Section IV.

We will discuss human-centric vision tasks and their status next, while highlighting the challenges and how our limited understanding of human brain functions (e.g. attention mechanisms, semantic models, knowledge-based reasoning) can be leveraged to effectively address some of the challenges.

III. HUMAN-CENTRIC VISION TASKS

Human understanding in the video is about the detection and recognition of humans, their attributes, and their activities. Significant progress has been made for many important human-centric vision tasks. We provide an overview of these technologies in the following.

A) People tracking

Visual object tracking is one of the fundamental problems in video analysis and understanding. Given the bounding box of a target object in the first frame of a video, a tracker is expected to locate the target object in all subsequent frames. Single object tracking can be typically considered a joint detection and tracking problem as it is essentially the detection of the *same* target object in every subsequent frame. The greatest challenge is to fulfill the simultaneous, but somewhat conflicting requirements on robustness and discrimination power [7]. The robustness requires a tracker not to lose tracking when the appearance of the target changes due to illumination, motion, view angle, or object deformation. Meanwhile, a tracker is expected to have the capability to discriminate the target object from a cluttered background or similar surrounding objects. Both requirements traditionally need to be handled through online training to achieve the adaptability.

Since 2015, more and more works based on CNN have emerged. While deep features introduce the speed limitation to online training, their strong representational power opens-up a possibility to completely remove online training. The pioneering work along this line is SiamFC [8]. SiamFC employs offline-trained Siamese CNNs to extract features, and then uses a simple cross-correlation layer to

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

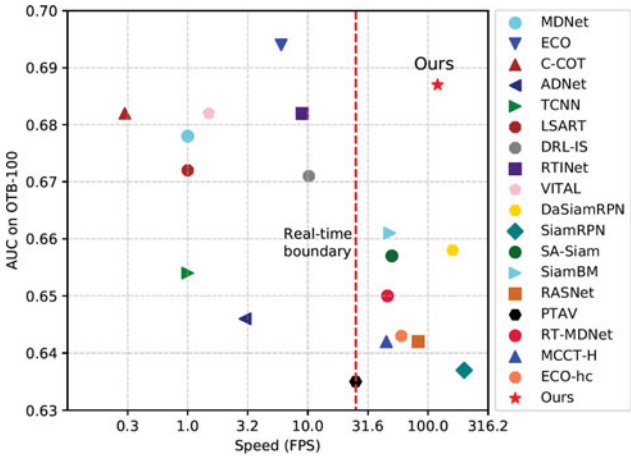

Fig. 2. Accuracy-speed trade-off of top-performing trackers on the OTB-100 benchmark. The speed axis is logarithmic. Reproduced from Fig. 8 of [7]. Please refer to [7] for the notations of different trackers. "Ours" refers to the SPM-Tracker [7].

perform a dense and efficient sliding-window-based search. As a result, SiamFC can operate at 86fps on GPU. There are a great number of follow-up works, such as SA-Siam [9] and SiamRPN [10]. SA-Siam adopts a two-branch network to encode images into two embedding spaces, one for semantic similarity (robustness) and the other for appearance similarity (discriminativeness). SiamRPN consists of a Siamese subnetwork for feature extraction and a region proposal subnetwork for similarity matching and box regression.

While most object tracking works use a single-stage design, a two-stage SiamFC-based network is proposed in [7], aiming to address both the robustness and discriminativeness requirements. The two stages are the coarse matching stage (i.e. a proposal stage) which focuses on enhancing the robustness, and the fine matching stage which focuses on improving the discrimination power by replacing the cross-correlation layer with a more powerful distance learning subnetwork. The resulting tracker achieves superior performance on major benchmark datasets, with an impressive inference speed of 120fps on an NVIDIA P100 GPU. Figure 2 shows an accuracy-speed trade-off of top-performing trackers on the OTB-100 benchmark [11]. It can be observed that significant progress has been made for single-object tracking in the past few years. These can be readily applied to single-person tracking.

In many practical scenarios, it is required to track multiple people simultaneously. It is generally not efficient to treat multiple-person tracking as multiple *separate* single-person tracking tasks. Recent works on multi-person tracking focus on the tracking-by-detection approach, i.e. detecting objects (of the same target category, i.e. person) using a general object/person detector in individual frames and then linking detections across frames. These include various strategies such as importance sampling and particle filtering for state propagation in a Bayesian framework [12], linking short tracks over a long duration, e.g. using the Hungarian algorithm for the optimal assignment [13], and greedy Dynamic Programming in which trajectories are estimated one after another [14]. To improve robustness to wrong identity assignment, recent research has focused on linking detections over a larger time duration using various optimization schemes. One common formulation to address multi-person tracking is based on constrained flow optimization which can be solved using the k-shortest paths algorithm [15]. Graph-based minimum cost multicut formulation has also been proposed [16, 17]. Note that the tracking-by-detection approach separates tracking from detection, therefore may not be the most efficient approach. For example, a general object detector may miss detecting some object instances in some frames, which otherwise could be tracked if a single-object tracker is used for that specific object instance. We believe more efforts should be made to investigate the approach of joint detection and tracking, which has been extensively studied for single-object tracking, for *joint* multiple-object/people tracking. The spatial and temporal relationships between multiple objects/people should be better exploited. There is also a trade-off between complexity and accuracy that should be optimized. Tracking over a long period of time is typically very challenging. Long-term tracking usually results in intermediate short-term tracklets, and requires linking/matching tracklets over time, e.g. through object re-identification (re-ID) techniques.

While general object tracking can be readily applied to people tracking, there are dedicated tracking technologies designed for people tracking. For example, a detector can be designed specifically for people to handle occlusion and body deformation [18]. A minimum cost-lifted multicut formulation was proposed in [19] to introduce additional edges in the graph to incorporate long-range person re-ID information into the tracking formulation. To effectively match hypotheses over longer temporal gaps, new deep architectures for re-ID of people are developed, where holistic deep feature representations and extracted body pose layout are combined. More discussion about person re-ID will be presented in Section III.C.

B) Human pose estimation

Human pose estimation determines the pixel locations of key joints of the human body. It is a key step toward understanding people. It has widespread applications such as human action recognition, motion analysis, activity analysis, and human–computer interaction. Despite many years of research with significant progress made recently, pose estimation remains a very challenging task, mainly due to the large variations in body postures, shapes, capturing views, scales, complex inter-dependency of parts, appearances, quality of images, etc.

The pictorial structure [20] is an early work that defines the deformable configurations by spring-like connections between pairs of parts to model complex joint relations. Subsequent works [21, 22] extend this idea to CNNs. Many recent works use CNNs to learn feature representations and obtain the locations of the 2D joints or the score maps of the

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

2D joints [23–25]. Some methods directly employ learned feature representations to regress the 2D joint positions [23], while a popular way of a joint detection is to estimate a score map for each 2D joint based on fully CNN [24–26]. To efficiently detect the 2D poses of multiple people, [26] uses Part Affinity Fields to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving real-time performance, irrespective of the number of people in the image. A two-stage normalization scheme, i.e. human body normalization followed by limb normalization, is presented in [27] to make the distribution of the relative joint locations compact, resulting in easier learning of convolutional spatial models and more accurate pose estimation.

In addition to 2D pose estimation discussed above, there have been recent efforts to estimate relative 3D poses from monocular images [28–30], mostly using regression methods. It is shown in [30] that a simple integral operation (replacing the "taking-maximum" operation by "taking-expectation") can relate and unify the heat map representation and joint regression, taking advantages of their individual merits to further improve the performance.

Another practically more important task is to estimate absolute 3D human poses from multiple calibrated cameras [31]. Many earlier works [32] follow the pipeline of first locating the 2D joints in each camera view and then triangulating them to 3D [31]. The 3D pose estimation accuracy heavily depends on the accuracy of the 2D joint estimations. In [33], a cross-view feature fusion approach to fuse the multi-view features in order to achieve more robust 2D pose estimation for each view, especially for the occluded joints, is presented. A Recursive Pictorial Structure Model (RPSM) is then presented to estimate 3D poses from multi-view 2D poses. This work sets a new state-of-the-art on multi-view human pose estimation on the benchmark H36M dataset [34], outperforming prior works (e.g. [35]) by a large margin (a 50% reduction of the average joint estimation error). This makes high accuracy motion capture without motion sensors or markers very close to reality, and is expected to enable many applications such as low-cost athlete body motion analysis, human tracking and action recognition in retail scenarios, etc.

C) Person re-identification

Person re-ID aims to match a specific person across multiple camera views or in different occasions from the same camera view. It facilitates many important applications, such as cross-camera tracking [36] and the long-term people tracking discussed in Section III.A. This task is very challenging due to large variations of person pose and viewpoint, imperfect person detection, cluttered background, occlusion, and lighting differences, etc. Many of these factors result in spatial misalignment of two matching human bodies, making it one of the key challenges in re-ID.

In recent years, many efforts have been made to alleviate these problems [37–39]. For example, to make the learned features focus on some local details, some works make a straightforward partition of the person image into a few fixed rigid parts (e.g. horizontal stripes) and learn detailed local features [37, 40, 41]. Such partition however cannot align well with the human body parts. Some other works have attempted to use the pose to help localize body parts for learning part-aligned features [38, 42, 43]. This, however, is a very coarse alignment. Even for the same type of parts, there is still spatial misalignment within the parts between images, where the human semantics are different for the same spatial positions. It becomes critical to design an architecture that enables the efficient learning of densely semantically aligned features for re-ID.

A densely semantically aligned person re-ID framework is proposed in [44], which enables fine-grained semantic alignment and explores the semantically aligned feature learning. It performs dense semantic alignment of the human body on a canonical space to address the misalignment challenges, where dense (i.e. pixel-wise) semantic pose estimation [45] is leveraged. To address the potential dense pose estimation errors (i.e. robustness issue) and the challenges in handling non-overlapping areas between two matching persons, a powerful joint learning framework is developed to guide the feature learning of one main stream using another (densely semantically aligned, but noisy) stream. State-of-the art performance is achieved on the benchmark datasets. Along the same line, a more elegant framework is proposed in [46] that uses an encoder–decoder architecture to guide the feature learning such that the learned features for re-ID is capable of reconstructing, through the decoder, a 3D full body texture image in a canonical semantic space. The learned features used for re-ID are thus view- and pose-invariant. The decoder is used only for model training, without increasing the inference complexity. This approach nicely addresses the issue of visible body inconsistency between matching images, which is not well-addressed in [44].

D) Human action recognition

Ultimately, we would like to understand the human activities in the video. Human action recognition thus is a very important but challenging task. There are a large variety of actions, with large or subtle differences. It is therefore important to be able to leverage some attention mechanisms, just like the human brain does, to focus on what really characterize a particular action to differentiate it from other actions. View variation (e.g. among videos taken of the same action) is another significant challenge, which demands view-invariant approaches.

There have been RGB-based approaches [47–53], skeleton-based approaches [54–57], and their combinations [58] for human action recognition. RGB-based approaches have the advantage of taking into account the appearance information and the context (e.g. background and other objects around humans), but have difficulty differentiating some fine-grained actions (e.g. human poses/motions). Skeleton-based approaches can focus on

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

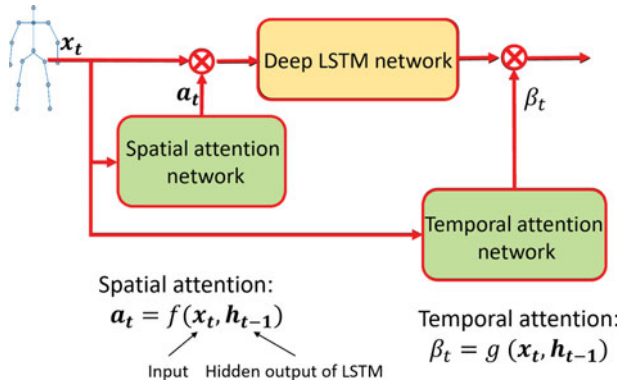

Fig. 3. Spatial-temporal attention network. Both attention networks use a one-layer LSTM network.

Fig. 4. Illustration of a retail intelligence scenario where multiple cameras are deployed, 3D space is reconstructed, people are detected and tracked, and heatmap (in purple) is generated.

fine-grained actions of the human body, but lose the context information (e.g. interacting objects). Combining the RGB information and pose information would often provide the best performance [58].

We focus on skeleton-based approaches here. Earlier neurophysiological study shows that human can recognize action by motions of just a few key points [59]. In fact, the skeleton can be considered as a high-level abstract representation of the human body, and it reflects some sort of human brain attention. Since there is a sequence of skeleton joint sets over time in the video, one can use a recurrent neural network such as Long Short-Term Memory (LSTM) network [60] to model the temporal dynamics of the joints for different actions. Co-occurrence patterns of joint features are learnt using an end-to-end regularized LSTM network for human action recognition [55]. Based on the observation that for different actions, the importance levels of different joints are different, both a spatial attention model and a temporal attention model are used to highlight what really characterize a particular action instance [56]. Figure 3 shows the network architecture. The single-layer LSTM-based spatial attention model assigns different weights to joints of the same frame before applying a baseline action classification network. The single-layer LSTM-based temporal attention model generates a temporal weighting curve to pool action prediction outputs at different time instances. Significant accuracy improvements (up to 6% absolute gains over baseline on a benchmark dataset) have been achieved, signifying the importance of developing attention mechanisms in vision tasks.

As mentioned above, a significant challenge for video analysis is view variation, i.e. visual representations of the same event captured from different views would look very different. The human brain has the capability to recognize that they are the same event. We would like the machine to be able to do that as well. A view adaptation sub-network is proposed in [57] to address the view-invariant property, by adaptively transforming the input 3D skeleton sequence to a more consistent virtual view before an LSTM-based classification network is applied. Through end-to-end training, this view adaptation sub-network allows the main action classification network to "see" only skeleton sequences of consistent views, disregard their original views. This effectively addresses the view variation problem, resulting in a powerful action classification network that can focus its attention on the action details, as opposed to visual content variation resulting from view variation.

E) Integration

While it is critical to develop individual human-centric vision tasks for video understanding, it is also important to look at them from a system perspective, and understand how to integrate those building blocks, and how they compensate each other.

Depending on the application scenarios, a practical system may integrate some or all of the building blocks. For example, for a workplace safety scenario, people detection and tracking may be sufficient to detect if there is any human activity outside a safety zone. In a more complex retail intelligence scenario where multiple cameras are deployed (see Fig. 4), one may need to detect and track the customers, based on face, body, skeleton, or a combination. Long-term tracking may be necessary and is typically challenging, especially across different cameras, in which case person re-ID becomes critical to link estimated short-term tracklets. With people tracking, heatmap can be created to reflect where in the store customers go/stay most. If more detailed activities of the customers are of interest, one may want to also identify the actions (e.g. picking up an item) of the customers, e.g. by leveraging the estimated pose sequence, or a combination of pose and RGB data. For efficient integration of the building blocks, a mask-RCNN [61] like architecture with a multi-task setting that shares the feature extraction backbone network could be used.

For real-time interactive applications such as video conferencing, the real-time requirement is likely the most significant challenge for many vision tasks. For example, foreground/human body segmentation and background blurring are desirable features to remove background distractions or provide privacy protection. In this case, there is

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

a stringent requirement on the neural network model size (e.g. in the order of 100 K bytes) and speed (in the order of ms per frame). Significant efforts need to be devoted to the model size reduction and speed optimization.

IV. FUTURE PERSPECTIVES

Understanding human in the video is a critical step for video understanding. There has been tremendous progress in the development of human understanding technologies in the past few years, thanks to the rapid advance of deep learning technologies. The development, however, has mostly focused on individual component vision tasks that are spatially or temporally local, such as detection of individual objects or actions. There is little work on understanding the relationship between individual entities, such as object relationships, and causal relations between actions/events. This is partly due to the difficulty in acquiring exponentially increasing amount of labeled data required by current deep learning technologies for complex tasks spanning over larger spatial regions and temporal durations. To address such difficulty, existing human knowledge should be leveraged and incorporated into the learning systems to efficiently learn about semantic relationships, while reducing the dependency on data-driven deep learning approaches. It is exciting to see that some initial efforts have been made along this line, where human knowledge is incorporated, e.g. in the form of graph convolutional network (GCN), an efficient variant of CNNs which operates directly on graphs [4], to help significantly improve the performance of video understanding [4–6]. For example, building upon GCN to transfer knowledge obtained from familiar classes to describe the unfamiliar class, [5] uses both semantic embeddings and the categorical relationships derived from a learned knowledge graph to predict the visual classifiers for unfamiliar classes. A Symbolic Graph Reasoning (SGR) layer, injected between convolution layers, is proposed in [6] to perform reasoning over a group of symbolic nodes whose outputs explicitly represent different properties of each semantic entity in a prior knowledge graph.

The development of semi-supervised learning or unsupervised learning technologies is also critical in alleviating the requirement on the amount of labeled training data, ultimately making it closer to mimicking how the human brain works. For example, recently unsupervised pre-training of deep Bidirectional Encoder Representations from Transformers (BERT), that is able to leverage the abundance of un-labelled data, shows a promising direction for language modeling [62] and joint visual-linguistic modeling [63]. The pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of downstream tasks, without substantial task-specific architecture modifications.

Although some aspects of human brain functions such as attention mechanisms have been well exploited in deep learning, it is still a long way to go in understanding better how human brain works before additional breakthroughs can be achieved in visual understanding. However, a human centric mindset, which includes both focusing on the human understanding tasks in video and leveraging our understanding of how the human brain works, will put us on the right track in developing video understanding technologies.

Although significant progress in research has been made for visual understanding, its landing in practice has been relatively slow. Industrial powerhouses and start-ups are rushing to push the technologies to the markets in different vertical domains, such as the Microsoft Cognitive Services (https://azure.microsoft.com/en-us/services/cognitive-services/). Face recognition is arguably the most successful (human understanding) vision technology that has found its widespread applications in practice. We are seeing more and more emerging real-world application scenarios such as retail intelligence, eldercare, workplace safety, and public security. Human understanding is a common core requirement in these scenarios, and we expect that the technologies will be mature in the near future to enable these application scenarios.

ACKNOWLEDGEMENT

The author would like to thank his colleagues and interns at Microsoft Research Asia for many discussions that helped shape the views and insights presented in this paper. Thanks especially go to Cuiling Lan, Chong Luo, Xiaoyan Sun, and Chunyu Wang for their long-term, close collaborations on the technical subjects discussed in the paper.

FINANCIAL SUPPORT

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

CONFLICT OF INTEREST

None.

REFERENCES

1 Ripley D.L.; Politzer T.: Introduction to the special issue – vision disturbance after TBI. *NeuroRehabilitation*, **27** (2010), 215–216. DOI: 10.3233/NRE-2010-0599.

2 He K.; Zhang X.; Ren S.; Sun J.: Deep residual learning for image recognition, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* Las Vegas, 2016.

3 Bahdanau D.; Cho K.; Bengio Y.: Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, 2014.

4 Kipf T.N.; Welling M.: Semi-supervised classification with graph convolutional networks, in *the Int. Conf. on Learning Representations,* San Juan, 2016.

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

5 Wang X.; Ye Y.; Gupta A.: Zero-shot recognition via semantic embeddings and knowledge graphs, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

6 Liang X.; Hu Z.; Zhang H.; Lin L.; Xing E.P.: Symbolic graph reasoning meets convolutions, *Advances in Neural Information Processing Systems*, Montreal, 2018.

7 Wang G.; Luo C.; Xiong Z.; Zeng W.: SPM-tracker: series-parallel matching for real-time visual object tracking, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, 2019.

8 Bertinetto L.; Valmadre J.; Henriques J.F.; Vedaldi A.; Torr P.H.: Fully-convolutional Siamese networks for object tracking, in *European Conf. on Computer Vision Workshop*, Springer, 2016, 850–865.

9 He A.; Luo C.; Tian X.; Zeng W.: A twofold Siamese network for real-time object tracking, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, 4834–4843.

10 Li B.; Yan J.; Wu W.; Zhu Z.; Hu X.: High performance visual tracking with Siamese region proposal network, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, 2018, 8971–8980.

11 Wu Y.; Lim J.; Yang M.-H.: Online object tracking: a benchmark, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, 2013, 2411–2418.

12 Giebel J.; Gavrila D.; Schnorr C.: A Bayesian framework for multi-cue 3D object tracking, in *European Conf. on Computer Vision*, Prague, 2004.

13 Perera A.; Srinivas C.; Hoogs A.; Brooksby G.; Wensheng H.: Multi-object tracking through simultaneous long occlusions and split-merge conditions, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, New York, June 2006, 666–673.

14 Fleuret F.; Berclaz J.; Lengagne R.; Fua P.: Multi-camera people tracking with a probabilistic occupancy map, *IEEE Trans. Pattern Anal. Mach. Intell.*, **2** (2008), 267–282.

15 Berclaz J.; Fleuret F.; Turetken E.; Fua P.: Multiple object tracking using k-shortest paths optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **9** (2011), 1806–1819.

16 Ristani E.; Tomasi C.: Tracking multiple people online and in real time, in *Asian Conf. on Computer Vision*, Singapore, 2014.

17 Tang S.; Andres B.; Andriluka M.; Schiele B.: Subgraph decomposition for multi-target tracking, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, 2015.

18 Tang S.; Andriluka M.; Milan A.; Schindler K.; Roth S.; Schiele B.: Learning people detectors for tracking in crowded scenes, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Sydney, 2013.

19 Tang S.; Andriluka M.; Andres B.; Schiele B.: Multiple people tracking by lifted multicut and person re-identification, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, 2017.

20 Yang Y.; Ramanan D.: Articulated pose estimation with flexible mixtures-of-parts, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, 2011.

21 Chen X.; Yuille A.: Articulated pose estimation by a graphical model with image dependent pairwise relations, in *Advances in Neural Information Processing Systems*, Montreal, 2014.

22 Yang W.; Ouyang W.; Li H.; Wang X.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, 2016.

23 Toshev A.; Szegedy C.: DeepPose: human pose estimation via deep neural networks, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, 2014.

24 Wei S.-E.; Ramakrishna V.; Kanade T.; Sheikh Y.: Convolutional pose machines, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, 2016.

25 Newell A.; Yang K.; Deng J.: Stacked hourglass networks for human pose estimation, in *European Conf. on Computer Vision*, Amsterdam, 2016.

26 Cao Z.; Simon T.; Wei S.; Sheikh Y.: Realtime multi-person 2D pose estimation using part affinity fields, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, 2017.

27 Sun K.; Lan C.; Xing J.; Wang J.; Zeng W.; Liu D.: Human pose estimation using global and local normalization, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, 2017.

28 Martinez J.; Hossain R.; Romero J.; Little J.J.: A simple yet effective baseline for 3D human pose estimation, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, 2017.

29 Moreno-Noguer F.: 3D human pose estimation from a single image via distance matrix regression, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, 2017, 1561–1570.

30 Sun X.; Xiao B.; Wei F.; Liang S.; Wei Y.: Integral human pose regression, in *European Conf. on Computer Vision*, Munich, 2018.

31 Hartley R.; Zisserman A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, 2003.

32 Amin S.; Andriluka M.; Rohrbach M.; Schiele B.: Multiview pictorial structures for 3D human pose estimation, in *British Machine Vision Conf.*, Bristol, 2013.

33 Qiu H.; Wang C.; Wang J.; Wang N.; Zeng W.: Cross view fusion for 3D human pose estimation, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Seoul, 2019.

34 Ionescu C.; Papava D.; Olaru V.; Sminchisescu C.: Human3. 6 m: large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Machine Intell.*, **7** (2014), 1325–1339.

35 Tome D.; Toso M.; Agapito L.; Russell C.: Rethinking pose in 3D: multi-stage refinement and recovery for markerless motion capture, in *Int. Conf. on 3D Vision*, Verona, 2018.

36 Wang X.: Intelligent multi-camera video surveillance: a review. *Pattern Recognit. Lett.*, **34** (1) (2013), 3–19.

37 Varior R.R.; Shuai B.; Lu J.; Xu D.; Wang G.: A Siamese long short-term memory architecture for human re-identification, in *European Conf. on Computer Vision*, Amsterdam, 2016.

38 Su C.; Li J.; Zhang S.; Xing J.; Gao W.; Tian Q.: Pose-driven deep convolutional model for person re-identification, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, 2017.

39 Suh Y.; Wang J.; Tang S.; Mei T.; Lee K.M.: Part-aligned bilinear representations for person re-identification, in *European Conf. on Computer Vision*, Munich, 2018.

40 Cheng D.; Gong Y.; Zhou S.; Wang J.; Zheng N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, 2016.

41 Wang G.; Yuan Y.; Chen X.; Li J.; Zhou X.: Learning discriminative features with multiple granularities for person re-identification, *ACM Multimedia*, Seoul, 2018.

42 Li D.; Chen X.; Zhang Z.; Huang K.: Learning deep context-aware features over body and latent parts for person re-identification, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, 2017.

43 Zhao H. *et al.* Spindle net: person re-identification with human body region guided feature decomposition and fusion, in *Proc. of the*

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press

IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, 2017.

44 Zhang Z.; Lan C.; Zeng W.; Chen Z.: Densely semantically aligned person re-identification, in *IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, 2019.

45 Guler R.A.; Neverova N.; Kokkinos I.: Densepose: dense human pose estimation in the wild, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

46 Jin X.; Lan C.; Zeng W.; Wei G.; Chen Z.: Semantics-aligned representation learning for person re-identification, in *AAAI Conference on Artificial Intelligence,* New York, 2020.

47 Weinland D.; Ronfard R.; Boyerc E.: A survey of vision-based methods for action representation, segmentation and recognition, *Comput. Vis. Image. Underst.*, **115** (2) (2011), 224–241.

48 Simonyan K.; Zisserman A.: Two-stream convolutional networks for action recognition in videos, in *Advances in Neural Information Processing Systems*, Montreal, 2014, 568–576.

49 Tran D.; Bourdev L.; Fergus R.; Torresani L.; Paluri M.: Learning spatiotemporal features with 3d convolutional networks, in the *IEEE Int. Conf. on Computer Vision*, Santiago, Chile, December 2015.

50 Feichtenhofer C.; Pinz A.; Wildes R.: Spatiotemporal residual networks for video action recognition, in *Advances in Neural Information Processing Systems*, Barcelona, 2016, 3468–3476.

51 Wang L. *et al.* Temporal segment networks: towards good practices for deep action recognition, in *European Conf. on Computer Vision*, Amsterdam, 2016, 20–36.

52 Qiu Z.; Yao T.; Mei T.: Learning spatio-temporal representation with pseudo-3d residual networks, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, 2017, 5533–5541.

53 Zhou Y.; Sun X.; Zha Z.; Zeng W.: MiCT: mixed 3D/2D convolutional tube for human action recognition, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

54 Du Y.; Wang W.; Wang L.: Hierarchical recurrent neural network for skeleton based action recognition, in *IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, 2015, 1110–1118.

55 Zhu W. *et al.* Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in *AAAI Conf. on Artificial Intelligence*, Phoenix, 2016.

56 Song S.; Lan C.; Xing J.; Zeng W.; Liu J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in *AAAI Conf. on Artificial Intelligence*, San Francisco, 2017.

57 Zhang P.; Lan C.; Xing J.; Zeng W.; Xue J.; Zheng N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, 2017.

58 Song S.; Lan C.; Xing J.; Zeng W.; Liu J.: Skeleton-indexed deep multi-modal feature learning for high performance human action recognition, in *IEEE Int. Conf. Multimedia and Expo*, San Diego, July 2018.

59 Johansson G.: Visual perception of biological motion and a model for its analysis. *Perception Psychophys* **14** (2) (1973), 201–211.

60 Graves A.: Supervised Sequence Labelling with Recurrent Neural Networks, Springer, Berlin, Heidelberg, 2012.

61 He K.; Gkioxari G.; Dollar P.; Girshick R.: Mask R-CNN, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, 2017.

62 Devlin J.; Chang M.; Lee K.; Toutanova K.: BERT: pre-training of deep bidirectional transformers for language understanding, in *Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), Minneapolis, USA, June 2019.

63 Sun C.; Myers A.; Vondrick C.; Murphy K.; Schmid C.: VideoBERT: a joint model for video and language representation learning, in *Proc. of the IEEE Int. Conf. on Computer Vision*, Seoul, 2019.

Wenjun Zeng received his B.E., M.S., and Ph.D. degrees from Tsinghua University in 1990, the University of Notre Dame in 1993, and Princeton University in 1997, respectively. He is currently a Sr. Principal Research Manager and a member of the Senior Leadership Team at Microsoft Research Asia. He has been leading the video analytics research powering the Microsoft Cognitive Services, Azure Media Analytics Services, and Office Media Experiences since 2014. He was with the Computer Science Department of University of Missouri (MU) from 2003 to 2016, most recently as a Full Professor. Prior to joining MU in 2003, he worked for PacketVideo Corp, San Diego, CA, USA; Sharp Labs of America, Camas, WA, USA; Bell Labs, Murray Hill, NJ, USA; and Panasonic Technology, Princeton, NJ, USA. His research interest includes mobile-cloud media computing, computer vision, and video analytics. He is a Fellow of the IEEE.

https://doi.org/10.1017/atsip.2019.26 Published online by Cambridge University Press