Machine Vision and Applications (2023) 34:120
https://doi.org/10.1007/s00138-023-01474-3

EDITORIAL

Guest editorial: special issue on human pose estimation and its applications

Wei Tang[1] · Zhou Ren[2] · Jingdong Wang[3]

Published online: 13 October 2023

1 Topic description

Estimating the human posture from an image or a video is a fundamental task in computer vision. It not only enhances other vision tasks like action recognition, person re-identification, and virtual try-on but also facilitates many real-world applications including robotics, healthcare, sports, and retail. An effective and efficient human pose estimation system can enable robots to acquire skills from demonstrations, support physical therapists in diagnosing and rehabilitating patients, assist sports analysts and coaches in tracking and training athletes, and empower retailers to establish employee-free stores.

Thanks to the advancement of deep learning and the availability of large-scale datasets, the performance of state-of-the-art human pose estimation methods has drastically improved in recent years. These approaches can accurately estimate postures in daily activities and sports. However, several challenges persist. For example, (1) it remains challenging to estimate postures that occur rarely or are entirely absent in the training data; (2) handling complex scenarios, such as crowded environments, motion blur, low-light conditions, and occlusions, is still a formidable task; (3) there is a growing need to develop efficient models that can estimate human poses in real-time or on low-power devices; (4) exploring novel applications for human pose estimation that can bring societal benefits or transform industries is an exciting avenue of research.

✉ Wei Tang
tangw@uic.edu

Zhou Ren
renzhou200622@gmail.com

Jingdong Wang
wangjingdong@outlook.com

[1] University of Illinois Chicago, Chicago, IL, USA

[2] Amazon AWS, Seattle, WA, USA

[3] Baidu, Beijing, China

2 Purpose of special issue

This special issue on *Human Pose Estimation and Its Applications* aims to provide a significant collection of original algorithms, theories, and applications to this field. It welcomed high-quality contributions with a focus on (but not limited to) the following topics:

- Single-person or multi-person human pose estimation
- 3D human pose estimation (from a single image, multiple views, or a video)
- Human pose tracking
- Weakly, semi, few-shot, or self-supervised human pose estimation
- Efficient human pose estimation
- Fairness, ethics, accountability, and transparency in human pose estimation
- Human pose estimation beyond normal adults, e.g., infants and people with disabilities
- Literature review/survey
- Datasets and annotations
- Applications of human pose estimation in robotics, augmented or virtual reality, healthcare, retail, sports, autonomous driving, human–robot interaction and collaboration, etc.

3 Explanation of submissions

We received a total of 20 submissions in response to our open call for papers. Each submission underwent a rigorous peer-review process, adhering to the journal's high standards. Finally, 12 manuscripts were accepted. These accepted papers have undergone major or minor revisions and satisfactorily addressed the concerns raised by the reviewers and editors. The remaining 8 manuscripts were transferred to a more suitable journal or special issue within the Springer portfolio. The acceptance rate is 60%.

4 Overview of accepted papers

The accepted papers cover a wide range of topics, from fundamental human pose models with improved accuracy and efficiency to applications in human behavior understanding and assessment, sign language recognition, and multi-camera calibration.

4.1 Robust human pose estimation

When estimating the human pose from a partial image of a person, humans can readily locate keypoints outside the image by referring to visual clues such as the body size. However, current methods do not consider those keypoints and focus only on the bounded area of a given image. Soonchan Park and Jinah Park propose a new approach, termed Position Puzzle Network and Augmentation, to locate human keypoints outside the bounding box. The Position Puzzle Network expands the spatial range of keypoint localization by refining the position and the size of the target's bounding box. The Position Puzzle Augmentation enables the keypoint detector to locate keypoints not only within, but also outside the input image. Experimental results demonstrate the substantial improvement offered by the proposed approach, with an average increase of 39.5% in mean Average Precision (mAP) and 30.5% in mean Average Recall (mAR) when compared to baseline keypoint detectors.

Yong Wei et al. present a novel top-down approach for instance-level human parsing, leveraging multi-task learning and crowded pose estimation. The method comprises several innovative components: a path attention feature pyramid to enhance the learning of robust multi-scale shared semantic features; an instance-agnostic human parsing module for learning body part segmentation and edge information; a Mask-RCNN-based crowded pose estimation module employing hierarchical association rules to acquire pose information; and a fusion strategy for integrating various semantic and instance features. This new approach has demonstrated superior performance compared to the majority of state-of-the-art methods when evaluated on two widely used human parsing datasets.

4.2 Human pose estimation on low-power devices

Current human pose estimation approaches tend to increase the complexity of a deep neural network for improved performance, making them inapplicable on edge devices and mobile terminals with limited computing power. Yanping Li et al. address this dilemma by designing a new lightweight basic block module. This module employs deep separable convolution and the reverse bottleneck layer to expedite network computations and trim down the overall model's parameters. Experiments conducted on the COCO and MPII datasets validate the effectiveness of this approach in optimizing human pose estimation for resource-constrained devices.

Theofanis Kalampokas et al. present a comprehensive review and benchmark analysis of human pose estimation methods in the context of Unmanned Aerial Vehicle (UAV) operations, with a specific focus on their algorithmic performance and efficiency. UAVs often operate under resource constraints to preserve battery power. The review encompasses a broad spectrum of research spanning the past 22 years. The benchmark evaluation scrutinizes the performance of 36 human pose estimation models using three well-established datasets. The results indicate that models based on the MobileNet architecture exhibit competitive performance while consuming fewer computational resources in comparison to ResNet-based models. Additionally, the benchmark findings are extrapolated to match the hardware specifications of edge devices, offering insights into the practicality of deploying these models in UAV applications.

4.3 Applications in human action recognition

Understanding human poses or skeletons is pivotal for action recognition. Many existing approaches rely on spatiotemporal graph convolutional networks (GCN). However, these methods often overlook interactions among features in the channel dimension, making it challenging to distinguish between actions with subtle differences. To address this limitation, Shuxi Wang et al. have introduced a novel Interactional Channel Excitation Module. This module dynamically recalibrates channel-wise pattern maps, enhancing the traditional GCN with channel-wise spatial excitation, channel-wise temporal excitation, and complementary graph topology. The proposed approach has surpassed previous state-of-the-art methods on three standard benchmark datasets, highlighting its superior performance in action recognition tasks.

Jiaji Liu et al. present a system that utilizes multi-person pose estimation in surveillance videos to enhance human behavior recognition. The process unfolds in three steps. First, multi-person pose estimation is carried out in the video using an end-to-end detection framework. Second, the pose estimates are refined by incorporating temporal cues among video frames. Finally, a human behavior recognition model, built upon the pose estimation, is deployed to identify classroom behaviors among students in video streams. Experimental results demonstrate that the human behavior recognition model, which integrates multi-person pose estimation and spatiotemporal semantics, yields significantly improved accuracy, showcasing its effectiveness.

4.4 Applications in human behavior assessment

Apraxia is an important symptom of Alzheimer's disease that manifests in its early stages. Traditionally, it is evaluated in clinical settings by having patients imitate hand gestures demonstrated by medical examiners. Cristina Vicedo et al. introduce a system designed to automate this assessment using videos of patients performing the gestures. The evaluation process comprises two main steps: (1) extraction of hand skeletons from the video footage and (2) employing a similarity function to quantitatively assess the accuracy of the gesture execution. The authors present experimental results involving several patients performing various gestures, demonstrating the effectiveness of the proposed method. This system aims to serve as a diagnostic tool, enabling medical professionals to identify potential mobility impairments in individuals with Alzheimer's disease.

Muhammad Ahmed Raza and Robert B. Fisher explore the potential of human pose estimation techniques in modeling and evaluating human eating behavior, which is of great importance for vision-based systems aimed at supporting independent living for the elderly. To achieve this goal, the authors design a state diagram that represents the most common eating behavior among subjects of all ages and an uncertainty-aware regression model that generalizes across different human subjects based on their distinctive micromovements. Additionally, the authors present an extended version of the EatSense dataset, which delves into eating behavior and the assessment of motion quality during eating activities.

Vaibhav Gulati et al. propose an automated approach for evaluating the soft skills of an individual via posture evaluation and vocal assessment. The evaluation integrates both non-verbal and verbal confidence scores of an individual. The non-verbal confidence score is estimated using the combination of a skeleton estimation algorithm and a posture angle evaluation system. The verbal confidence score is estimated using pause detection, filler word detection, and continuous word repetition estimation models. Experimental results validate the effectiveness of this new approach. Notably, the proposed system can help users improve the quality of their presentation skills, personality, and employability.

4.5 Applications in sign language recognition

Research in the domain of sign language recognition holds the potential to bridge communication gaps between the deaf and able people. Nevertheless, much of the existing work in this field primarily concentrates on American Sign Language and Chinese Sign Language, with limited attention given to Pakistan Sign Language. A key challenge in this context is the scarcity of available data, making it extremely challenging to effectively train deep neural networks. Hafiz Muhammad Hamza and Aamir Wali introduce a pipeline for a Pakistan Sign Language recognition system that incorporates a data augmentation unit. To evaluate the effectiveness of this pipeline, three deep learning models—C3D, I3D, and TSM—are employed. Among these models, C3D emerges as the most suitable choice, achieving an accuracy rate of 93.33% while also enjoying a shorter training time compared to the other models.

Despite recent advancements, sign language recognition research remains fragmented, often limited to very small vocabularies, and primarily tailored to specific conditions. Aamir Wali et al. offer a comprehensive survey of state-of-the-art sign language recognition techniques developed between 2019 and 2022. The review delves into a wide array of frameworks based on different linguistic units, including alphabets, words, and sentences. It covers various models, ranging from convolutional neural networks to Transformer-based approaches. In addition, this review provides a summary and comparison of datasets that have been widely used in recent years.

4.6 Applications in multi-camera calibration

To precisely reconstruct an object or person in the 3D space, a common pipeline is to first set up multi-view cameras in the target area and then perform triangulation of the matching joints to calculate the 3D coordinates. However, calibrating such a setup typically requires dedicated equipment and elaborated test procedures. S. Dehaeck et al. introduce a new calibration method based only on the detection of one or more people walking through the field of view. The new method allows the calibration to happen simultaneously with the measurements being taken, which is practical when dealing with uncontrolled environments. Experimental results show that this calibration procedure is more accurate than a typical incremental calibration procedure using a chessboard.

5 Conclusions

This special issue presents a collection of high-quality papers on human pose estimation and its applications. They cover a wide range of topics, from fundamental human pose models with improved accuracy and efficiency to applications in human behavior understanding and assessment, sign language recognition, and multi-camera calibration. Through this special issue, we would like to open the path toward more discussion between practitioners and researchers in this very important yet challenging field of human pose estimation.

Wei Tang is currently an Assistant Professor in the Department of Computer Science at the University of Illinois Chicago. He obtained his Ph.D. from Northwestern University. His research interests are computer vision, pattern recognition, and machine learning. He is an Associate Editor of IAPR Machine Vision and Applications, and IET Computer Vision. He served as an Area Chair for CVPR, WACV, ICPR, and FG.

Zhou Ren is currently an Applied Science Manager at Amazon AWS. He was a Principal Research Manager and founding member at Wormpex AI Research from 2018 to 2022. Before that, he was a Senior Research Scientist at Snap Inc. from 2016 to 2018. He received his Ph.D. from the Computer Science Department at University of California Los Angeles. Zhou Ren is currently an Associate Editor of The Visual Computer Journal and Chair of Industrial Publication Committee at Asia-Pacific Signal and Information Processing Association. He served as an Area Chair for CVPR 2021, CVPR 2022, WACV 2022, WACV 2023, WACV 2024, and as a Senior Program Committee for AAAI 2021, AAAI 2022. His current research interests include computer vision, video analysis, and multimodal understanding, etc. Zhou Ren is the recipient of the IEEE Transactions on Multimedia 2016 Best Paper Award and he is the CVPR 2017 Best Student Paper Award nominee. He has won the runner-up prize of the NIPS 2017 Adversarial Attack and Defense Competition and the 1st Prize in ICCV 2021 Low Power Computer Vision Challenge.

Jingdong Wang is Chief Scientist for Computer Vision with Baidu. Before joining Baidu, he was a Senior Principal Research Manager with Microsoft Research Asia. His areas of interest are computer vision, deep learning, and multimedia search. His representative works include deep high-resolution network (HRNet), object-contextual representations for semantic segmentation (OCRNet), neighborhood graph search (SPTAG) for large-scale vector search. He has been serving/served as an Associate Editor of IEEE TPAMI, IJCV, ACM TOMM, IEEE TMM, and IEEE TCSVT, and an area chair of leading conferences in vision, multimedia, and AI, such as CVPR, ICCV, NeurIPS, ECCV, ACM MM, IJCAI, and AAAI. He was elected as an ACM Distinguished Member, a Fellow of IAPR, and a Fellow of IEEE, for his contributions to visual content understanding and retrieval.