Human Pose Estimation: Single-Person and Multi-Person Approaches

Wan Tang

School of Mathematics and Statistics, South-Central Minzu University, Hubei Province, China

Abstract. Human pose estimation (HPE), as one of the core tasks in computer vision, plays a crucial role in enabling computers to comprehend human behaviour interactions. With the advancement of technology, this task has demonstrated significant potential in various application areas such as motion capture, behavior analysis and augmented reality. Despite significant progress in recent years, HPE still presents challenges when dealing with complex scenarios such as occlusion, illumination changes, and dynamic backgrounds. This paper will provide a comprehensive overview of HPE techniques, focusing on both single-person and multi-person poses. According to their respective characteristics and application scenarios, single-person pose estimation is categorized into traditional methods and deep learning methods, while multi-person pose estimation is classified into top-down and bottom-up aspects. In addition, this paper analyzes commonly used datasets relevant to HPE, discusses the current unsolved issues, and forecasts future research directions, with the aim of providing valuable references and guidance for subsequent research in this field.

1 Introduction

Human pose estimation (HPE) is a crucial computer vision technology that aims to infer human pose by identifying and localizing key points on the human body. This technique has demonstrated great application potential in the fields of motion capture [1], behavior analysis [2], sports training [3], augmented reality [4], and virtual reality [5], and is crucial for computers to understand human movements. In the previous few years, with the accelerated growth of deep learning techniques, particularly the innovations in convolutional neural networks (CNNs), HPE has shown significant improvement in accuracy and robustness. This paper will provide an overview of research advances in the field and discuss current challenges and future solutions for the field. This paper aims to provide a comprehensive perspective for new researchers to help them gain a deep understanding and keep up with research developments.

This paper reviews and analyzes the research methods in the field of HPE from both single-person and multi-person perspectives. This research will first analyze the application strategies of traditional methods and deep learning methods in single-person pose estimation respectively. Subsequently, the benefits and drawbacks of both top-down and bottom-up

Corresponding author: 02221101147@mail.scuec.edu.cn

approaches to multi-person pose estimation are contrasted. In addition, this paper will outline the commonly used datasets in the HPE domain, showing the performance of different algorithms applied to each dataset. Finally, this paper analyzes the current limitations of HPE techniques and provides an outlook on their future directions and potential improvements, aiming to provide guidance and insights for continuous development and innovation in the field. Through this structural arrangement, this paper not only provides readers with a comprehensive overview of the HPE field but also points out the direction for future research exploration.

2 Human pose estimation methods

2.1 Single-person pose estimation

With the primary objective of identifying the important parts of the human body, single-person pose estimation seeks to identify and estimate the human body stance of a single person from an image or video. The two primary categories of single-person pose estimation techniques are traditional techniques and deep learning-based techniques.

2.1.1 Traditional methods

Traditional methods are principally based on the graph structure model, pictorial structures model (PSM) [6, 7] is a graph-based representation for representing human body poses. The model represents the human body as a graph structure, with nodes standing in for key parts of the body and edges for spatial relationships between these parts. This representation effectively captures the structure and spatial relationships of the human body and provides an intuitive and structured way to perform pose estimation. In graph-structured models, constructing and optimizing the graph model usually requires significant computational resources. Specifically, the construction of the model requires determining the locations of the nodes by matching the image features of the body parts and also adjusting the weights of the edges by optimizing the spatial relationships. This process involves complex computational and optimization algorithms to ensure that the model accurately captures human body poses. In addition, feature extraction relies on manually designed algorithms. These algorithms typically extract information from images based on specific visual features. However, the performance of manually designed feature extraction algorithms tends to vary widely in different scenes and conditions, limiting the generalization and flexibility of the model. In other words, these methods may have difficulty adapting to different lighting, occlusion, and pose changes when dealing with diverse and complex pose estimation tasks. Nonetheless, graph-structured models played an important role in early pose estimation research. They laid the foundation for subsequent deep learning approaches by providing a structured representation that enables researchers to better understand and model complex pose relationships in the human body. However, deep neural network-based techniques have rapidly supplanted graph-structured model-based approaches due to the speedy advancement of techniques for deep learning. These approaches have shown to be more capable and flexible in handling complicated scenarios and enormous amounts of data.

2.1.2 Deep learning methods

Pose estimation techniques based on deep learning have gained popularity as these techniques have advanced. The application of deep CNNs achieves automation and high efficiency in feature extraction and is able to extract richer and more robust feature representations from

images. Toshev et al. proposed deep pose estimation (DeepPose) [8] in 2014, which implements HPE through deep neural networks. The method uses a cascaded CNN to extract features from the original RGB image and gradually refines the prediction of the joints by regression technique to achieve accurate pose estimation. Convolutional pose machines (CPM), a multi-stage CNN, was proposed by Wei et al. in 2016 [9]. The backbone network for feature extraction is the visual geometry group network (VGGNet). By adding supervisory signals at each stage and implicitly learning the spatial relationships between joints, CPM improves key point position prediction and increases the model's robustness in complicated settings. In the same year, Newell et al. proposed stacked hourglass networks (SHNs), which uses several hourglass-shaped network structures stacked together, with each module forming a symmetric U-shaped structure through downsampling and upsampling [10]. By jumping connections, this design tackles the issue of information loss in multi-scale feature extraction and enables the network to capture the spatial relationship between human joints at various resolutions. In 2019, in order to maintain high-resolution spatial information and prevent the loss of detailed information caused by downsampling in the traditional network structure, the high-resolution network (HRNet) was proposed [11]. HRNet is a network that is parallelized through a multi-branching structure and feature fusion strategy. This enables the network to use high-resolution features for more accurate pose capture at each stage.

2.2 Multi-person pose estimation

The goal of multi-person pose estimation is to identify, locate, and estimate each person's body keypoints inside an image or video while also differentiating between each person's keypoints. The field is mainly categorized into top-down and bottom-up approaches.

2.2.1 Top-down

The top-down approach recognizes every person in the picture first, and then it estimates each person's unique pose. This method performs well in distinguishing different individuals but may encounter difficulties in detection and differentiation when dealing with figure occlusion or high pose similarity. Mask region-based convolutional neural network(R-CNN) is another approach and it achieves pixel-level accurate segmentation by introducing a mask branch in the region-of-interest detection [12]. This method uses a feature pyramid to combine high-level semantic information with low-level specifics, especially in human key point detection, and its use of the ResNet-50-FPN backbone shows excellent performance. A hierarchical network model is designed using the cascaded pyramid network (CPN) to address complicated backdrop issues and occlusion in multi-person pose estimation [13]. Using a human body detector, CPN first determines the bounding box. Next, it uses the GlobalNet and RefineNet sub-networks to detect and refine the keypoints within the bounding box. GlobalNet is responsible for quickly identifying obvious keypoints, while RefineNet focuses on correcting occluded or difficult-to-identify keypoints. CPN Soft-NMS technique is used to improve the recall in the case of multiple overlapping people and improve the accuracy and robustness in crowded situations. The problem of pose estimation in congested scenarios is the focus of LI et al.'s 2019 CrowdPose model [14]. The model combines a joint single-person pose estimator with a human detector to locate key points within a bounding box, and utilizes a graph model for multi-peak prediction and global correlation. CrowdPose exhibits robustness in complex environments and guarantees efficient inference speed, making it particularly suitable for pose estimation in dense crowds.

2.2.2 Bottom-up

The bottom-up approach initially identifies every potential human keypoint in a picture, and then applies a priori knowledge of human structure to group keypoints and associate them with different individuals. This method shows advantages in multi-person pose estimation, known for its uniform detection of all keypoints in the image and efficient grouping associations based on a priori knowledge of human body structure, but it may encounter the problem of misconnecting keypoints when dealing with occlusion and complex scenes. OpenPose uses a cascaded CNN to extract features and generates part confidence maps (PCMs) and part affinity fields (PAFs) via a cyclic iterative network for precise location and association of keypoints [15]. OpenPose connects keypoints detected by PCMs via PAFs and optimizes the keypoints' connection by using the maximum-weighted bipartite map matching and Hungarian algorithm process. Newell et al. in 2017 proposed the Associative Embedding method [16], which transforms pose estimation into a clustering problem by assigning an embedded label to each detected keypoint to identify the group to which it belongs and groups the keypoints by comparing the similarity of the labels. Papandreou et al. 2018 proposed the PersonLab method [17], which combines keypoint detection and instance segmentation. It uses a ResNet network to anticipate the joint heatmap of human keypoints, the relative displacement of keypoints, and the human segmentation mask and groups the keypoints by a greedy decoding algorithm to realize the association with human instances. PersonLab demonstrates efficient pose estimation in the common objects in context (COCO) instance segmentation task. Kreiss et al. proposed part intensity fields and part association fields (PifPaf) in 2019 [18], the method is designed for low-resolution and crowded scenes. It employs a dual-head neural network to predict joint confidence, position and size, and inter-joint associations, respectively. PifPaf determines the keypoints by greedy decoding and optimizes using Laplacian and smoothing losses, demonstrating superior performance in complex scenes. Cheng et al. proposed HigherHRNet [19] in 2020 which intends to increase the precision of small-scale HPE. The method adds a transposed convolution module to HRNet to produce high-resolution heatmaps that are scale-aware and adopts a multi-resolution supervised training strategy to enable features at different levels to learn information at different scales.

3 Datasets

The commonly used datasets in HPE are shown in Table 1, where the leeds sports pose (LSP) dataset [20] contains 2000 images of sports scenes, focuses on single-person pose estimation, and labels 14 joints. The frames labeled in cinema (FLIC) dataset contains images from Hollywood movies, focuses on multi-person pose estimation, and labels 10 upper body joints [21]. The max planck institute for informatics human pose (MPII) dataset contains about 25K images of daily activities for single-person pose estimation, labeled with up to 16 body joints [22]. The COCO dataset is a large-scale object detection, segmentation, and pose estimation dataset containing 200K images for single and multi-person pose estimation, labeled with 17 keypoints [23]. Human3.6M dataset is a large-scale 3D human motion capture dataset containing four high-resolution human motion capture images and it contains multiple movements of 11 actors captured by 4 high-resolution cameras, providing accurate 3D joint positions and corresponding 2D images [24]. Posetrack multi-person video pose (PoseTrack) dataset is a video pose estimation dataset containing multiple video sequences focusing on multi-person pose estimation and tracking [25].

Table 1. Common datasets for HPE

Dataset	Year	Total	Train	Val	Test	Single-/multi-person	Keypoint
LSP	2010	2k images	1k	-	1k	Single-person	14
FLIC	2013	5k images	4k	-	1k	Single-person	10
MPII	2014	25k images	15k	3k	7k	Single-person	16
COCO	2014	108k images	64k	3k	41k	Single-person & multi-person	17
Human3.6M	2014	5k videos	3k	500	1.5k	Single-person	32
PoseTrack	2017	514 videos	250	50	214	Multi-person	15

From DeepPose [8] to HigherHRNet [19], deep learning models have significantly improved the accuracy of single-person pose estimation by automating feature extraction and multi-stage prediction. For example, the high accuracy achieved by HigherHRNet [19] on the MPII dataset exemplifies the advantages of deep learning when dealing with high-resolution features. Top-down methods such as Mask R-CNN [12] and CPN [13] achieve high mAP accuracy on the COCO through accurate human detection and keypoint localization. These methods enhance robustness in occlusion and crowded scenes through techniques such as feature pyramid and Soft Non-Maximum Suppression. Integrating additional human pose data, such as combining other data sources on the MPII dataset, can significantly improve the accuracy from 92% to over 94%. This finding suggests that the diversity and richness of datasets are crucial for improving the generalization ability and performance of deep learning models. Although bottom-up methods have an advantage in dealing with key points of multiple people in a single graph, top-down methods usually have higher accuracy in multi-person pose estimation. This may be because top-down methods simplify the problem's complexity through the strategy of detecting before estimating.

4 Challenges and future research directions

The field of HPE, despite significant technological advances, still faces many challenges. Among them, the occlusion problem is one of the major challenges in the field, and although bottom-up approaches such as BalanceHRNet [26] have gained some improvements by augmenting the sensory field of the network and some progress in multi-resolution representation, the robustness of the existing algorithms still needs to be improved when dealing with extreme occlusion and multiple overlapping human scenes. Future research may focus on utilizing deeper attention mechanisms and graph neural networks to strengthen the algorithm's occlusion handling capabilities, while the use of multiple-task learning is expected to further enhance the model's generalization and robustness. The problem of dataset diversity and imbalance also poses a challenge to the algorithm's generalization ability. Existing datasets may have data imbalance on certain poses or scenarios, which affects the algorithm's capacity for generalization. To solve this issue, data augmentation [27] and transfer learning techniques have been used to increase data diversity and enhance the models' capacity for generalization, but these methods may not fully address the data imbalance problem. Generative Adversarial Networks (GANs) and multi-task joint learning may be future solutions to improve model generalization ability. Real-time performance is another concern, especially in applications that require real-time feedback, such as augmented reality and virtual reality. Lightweight models [28] and model pruning techniques have been used to optimize the use of computational resources, but it is still a challenge to accomplish high precision in real-time performance. Network Architecture Search and hardware acceleration techniques may provide new ideas for real-time performance optimization in the future. Finally, the issue of model interpretability is also receiving increasing attention. Deep learning models are often considered black boxes and lack interpretability. Model visualization and post-processing interpretation techniques such as

feature attribution analysis have been used to improve model interpretability, but these methods often rely on specific models or datasets and lack generality. In the future, the development of interpretable algorithms and the construction of cross-model interpretation frameworks may provide new ways to improve model interpretability.

5 Conclusion

This review provides an in-depth look at current developments within the HPE domain. Despite the progress made in several aspects of current HPE techniques, there are still some limitations, such as the performance of existing algorithms in dealing with occlusion, lighting changes, and complex backgrounds need to be improved. This paper focuses on the key technological breakthroughs and innovative algorithms for single- and multi-person pose estimation. Firstly, from the early traditional approaches to the contemporary deep-learning based approaches, this study carefully analyzes the key algorithms for single-person pose estimation. Subsequently, multi-person pose estimation techniques, including two popular approaches, top-down and bottom-up, are combed in detail, and their respective advantages and limitations are analyzed. In addition, this paper introduces the commonly used datasets within the HPE domain and compares the performance and scores of different algorithms under these criteria. In the outlook section, the current unsolved problems in the field are analyzed and possible solutions are proposed. The review of the field provides researchers with an overview of the development of the HPE field, and provides a referable direction for solving existing problems and promoting future research, to provide more reference value for future research and promote the development of the field.

References

1. H.-S. Fang, J. Li, H. Tang, et al. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 7157-7173 (2022).
2. Y. Huang, S. Zhai, Y. Liu, Deep Learning Based a Novel Method of Classroom Behavior Recognition, in Proceedings of the 2022 IEEE 2nd International Conference on Educational Technology (ICET), Beijing, (2022).
3. J. Hwang, S. Park, N. Kwak, Athlete Pose Estimation by a Global-Local Network, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, (2017).
4. L. Zhu, Y. Chen, Y. Lin, C. Lin, A. Yuille, Recursive segmentation and recognition templates for image parsing, in proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, (2011).
5. Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2021).
6. F. Wang, Y. Li, Beyond physical connections: Tree models in human pose estimation, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2013).
7. L. He, G. Wang, Q. Liao, J.-H. Xue, Latent variable pictorial structure for human pose estimation on depth images. Neurocomputing. **203**, 52-61 (2016).
8. A. Toshev, C. Szegedy, DeepPose: Human Pose Estimation via Deep Neural Networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2014).

9. S.-E. Wei, V. Ramakrishna, T.-K. Kanade, Y. Sheikh, Convolutional pose machines, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, (2016).
10. A. Newell, K. Yang, J. Deng. Stacked hourglass networks for human pose estimation. in proceedings of the 14th European Conference on Computer Vision. (2016).
11. K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2019).
12. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in Proceedings of the IEEE international conference on computer vision, (2017).
13. Y. Chen, Z. Wang, Y. Peng, et al. Cascaded pyramid network for multi-person pose estimation, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2018).
14. J. Li, C. Wang, H. Zhu, et al. CrowdPose: efficient crowded scenes pose estimation and a new benchmark, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2019).
15. Z. Cao, T. Simon, S.-E. Wei, et al. Realtime multi-person 2D pose estimation using part affinity fields, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2017).
16. A. Newell, Z. Huang, J. Deng. Associative embedding: end-to-end learning for joint detection and grouping. in proceedings of the Conference on Advances in Neural Information Processing Systems, (2017).
17. G. Papandreou, T. Zhu, L.-C. Chen, et al. Personlab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, in Proceedings of the European conference on computer vision (ECCV), (2018).
18. S. Kreiss, L. Bertoni, A. Alahi, PifPaf: composite fields for human pose estimation, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2019).
19. B. Cheng, B. Xiao, J. Wang, et al. HigherHRNet: scale aware representation learning for bottom-up human pose estimation, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, (2020).
20. S. Johnson, M. Everingham, Clustered pose and nonlinear appearance models for human pose estimation. in Proceedings of the British Machine Vision Conference (BMVC), (2010).
21. B. Sapp, B. Taskar, Modec: Multimodal decomposable models for human pose estimation. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2013).
22. M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: New benchmark and state of the art analysis, in Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, (2014).
23. T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, Microsoft COCO: common objects in context, in European Conference on Computer Vision, 740–755(2014).
24. C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, in Proceedings of the IEEE transactions on pattern analysis and machine intelligence, (2013).

25. U. Iqbal, M. Garbade, J. Gall, Pose for action-action for pose. In proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), (2017).
26. Y. Li, S. Jia, Q. Li, BalanceHRNet: An effective network for bottom-up human pose estimation, Neural Netw. **161**, 297-305 (2023).
27. S.-S. Park, H.-J. Kwon, J.-W. Baek, K. Chung, Dimensional Expansion and Time-Series Data Augmentation Policy for Skeleton-Based Pose Estimation. IEEE Access. **10**, 112261-112272 (2022).
28. X. Zhang, X. Zhou, M. Lin, J. Sun, ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices, in Proceedings of the IEEE conference on computer vision and pattern recognition, (2018).