Multi-Person Pose Estimation: Method Classification and Cross-Dataset Performance Analysis

Zikun Li

School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
2023050906023@std.uestc.edu.cn

Abstract. Finding the important features of each human body in the picture and accurately allocating those features to each individual is the main challenge of multi-person posture estimation. Multi person pose estimation tasks can provide support for multiple downstream tasks and overcome the limitation of single person pose estimation that can only recognize a single human body. As deep learning has advanced in the field of machine vision, deep learning-based multi-person pose estimation techniques have progressively supplanted conventional techniques and are now widely used. The task is divided into two methods based on human body modeling: heatmap and vector field estimation. Among them, the heatmap method is further divided into top-down and bottom-up methods based on modeling logic. This article first divides the task of multi-person pose estimation into heatmap and vector field regression, and analyzes their characteristics by category. It then compares their performance on two datasets. Finally, it offers a prospect for future development. This article summarizes the advantages, disadvantages and main improvement directions of different types of implementation methods. Its comparison study also identifies the approaches' performance focus, which can serve as a guide for further research.

Keywords: Multi-person pose estimation, Deep learning, Top-down, Bottom-up, Vector field regression.

1 Introduction

The task of multi-person pose estimation (MPPE) is to predict the key point coordinates of each human body in the image obtained from a monocular camera. MPPE intends to simultaneously mark multiple human key points such as head, elbow, leg joint, foot, etc, and infer which ones belong to the same person, even if there is occlusion between different instances. This feature compensates for the limitation of single-person pose estimation(SPPE), which can only predict a single human body in one image. MPPE can be used to assist multiple downstream tasks, such as re-identification (Re-id) [1], action recognition [2], etc. MMPE is more convenient and economical compared to traditional human modeling methods because it does not require additional motion capture equipment. This technology has broad application scenarios in game production

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_16

and movie motion capture, and can also be applied to motion sensing games and augmented reality technology.

In early research, researchers typically used traditional machine learning methods to achieve this task. They extract features by designing shallow feature extractors and use them to locate key points. Although they have managed some achievements, these methods still cannot meet the accuracy requirements of application scenarios, and the complex design process also hinders the widespread application of traditional methods. As convolutional neural networks and deep learning become more sophisticated, the performance of feature extractors composed of deep neural networks far exceeds that of traditional manually designed feature extractors. Therefore, the current mainstream methods use neural networks to learn features and then locate key points through different strategies.

This article summarizes recent multi person pose estimation works, analyzes the performance of different types of methods, and looks forward to and summarizes the development prospects of multi person pose estimation tasks. The overall structure of the article is as follows. The first chapter briefly introduces the background of the task in the form of an introduction. The second chapter summarizes recent work according to the category of method implementation. The third chapter compares and analyzes the performance of typical methods. The fourth chapter provides an outlook on the task of multi person pose estimation. The fifth chapter is a conclusion.

2 Multi-Person Pose Estimation Classification

The task of MPPE can be accomplished through two modeling methods: heatmap and vector field regression to locate key points. The method of heatmap outputs a probability map consisting of Gaussian kernels representing key points, while vector field regression locates key points by outputting the regression vectors of the key points. There are two methods in the heat map: top-down and bottom-up.

2.1 Output Heat map

Top-down Logic. In the top-down approach, the algorithm first extracts individual pedestrians through a human body detector, and then uses mature single person pose estimation algorithms to estimate the limb key points of each pedestrian. This estimation method poses a significant performance challenge to human detectors, as SPPE is highly sensitive to local errors in bounding boxes and needs to keep the center of the human body at the center of the bounding box.

To address this issue, Fang et al. proposed a Symmetric Spatial Transformer Network (SSTN) to ensure the quality of the bounding box, while using Parameterized Posture Non Maximum Suppression (p-Pose NMS) to reduce redundant pose detection [3]. Afterwards, Sun et al. [4] proposed a backbone network that utilizes parallel multi-resolution subnetworks to achieve high-resolution output while having excellent multi-scale feature extraction capabilities, while "ViTPose" [5] uses a non-layered visual Transformer as the encoder, greatly improving the model's feature extraction ability. The above methods all require a separate detector, and in addition to separating the

model, there are also some methods that choose to integrate keypoint detection into the pedestrian detection model. "Mask-RCNN" improves on fast-RCNN by adding a branch for locating key points after the original network [6]. In this network, two tasks share a feature layer, but recognizing the human body requires more semantic information and key points require more spatial information. The shared feature layer leads to contradictions between these two tasks during the optimization process. Mao et al use FPN networks to extract features from different layers to correspond to different tasks, which to some extent solves the problem of optimization contradictions [7].

Bottom up Logic. In the bottom-up approach, the model first predicts all key points in the image, and then uses correlation methods to link the key points of the same instance together. Pishchulin et al turn the association process into an integer linear regression problem and use image conditioned pairwise terms to eliminate duplicate nodes. The use of integer linear programming cannot effectively handle the intersection of limbs in images [8]. Therefore, Cao et al propose using part affinity fields (PAFs) to match limbs, which can better connect intersecting limbs due to their feature of containing limb directions [9]. And “Associative Embedding” (AE) chooses to use the global features of key points to output a feature value that makes the key point feature values of the same instance similar, so as to be assigned to the same instance [10]. Cheng et al utilize the network of “HRNet” to improve the multi-scale feature extraction ability of 'AE' [11]. Luo et al. presented scale-adaptive heatmap regression to tackle the intrinsic uncertainty of key points and improved the accuracy of key point localization by adjusting the Gaussian kernel standard deviation [12].

2.2 Vector Field Regression

Vector field regression utilizes the offset of the network output relative to the center of the human body to locate key points. There are three branches in the Duan et al designed a highly anticipated vector field regression network called the “Centernet” [13]. This network outputs a human center point heatmap, key point offset values, and key point heatmap, and determines the position of the human body through the human center point heatmap. The closest key point to the regression position is selected as the most likely true key point. The birth of Centernet has aroused researchers' interest in vector field regression. However, due to the fact that Centernet only regresses key points based on the center point, the feature discrimination of key points is insufficient. At the same time, the network focuses on different feature scales when optimizing regression and heatmap output simultaneously, which makes optimization difficult. Therefore, "DEKR" proposed multi branch and adaptive convolution to extract different key point features, improving regression ability [14]. At the same time, directly regressing the absolute coordinates of key points solves the problem of unclear human center and optimization contradiction.

SPM did not choose to directly regress key point coordinates, but instead based on hierarchical SPR human key point modeling to predict from the head key points, and then predict the offset value for the previous node step by step[15]. This short-range and inexpensive encoding makes the prediction results closer to the human skeleton.

3 Methods for Performance Evaluation and Analysis

3.1 Dataset for Comparison

This study selected typical methods from each category to compare and analyze their performance on the COCO dataset and Crowdpose. The selection of the COCO dataset in this article is mainly due to its wide application, and most mainstream algorithms focus on performance on the COCO test set as the main indicator.

The main reason for choosing the Crowdpose dataset is that it has a higher human density and a more complex environment, making it suitable for comparing the performance of different methods in complex scenes.

The comparison indicators mainly use the average precision（AP） under different Object Keypoint Similarity (OKS) thresholds. In order to compare the average precision（AP） of various typical methods, this paper collected and organized the data of each typical method on the COCO2017 and CrowdPose datasets, as shown in Table 1 and Table 2.

Table 1. Performance Comparison of Single Scale Testing on COCO Test Set.

(The methods marked with * were trained using a multi-dataset)

method	backbone	AP	AP50	AP75	APM	APL
heatmap						
top-down						
HRNet[11]	HRNet-W48	75.5	92.5	83.3	71.9	81.5
TokenPose[16]	HRNet-W48	75.9	92.3	83.4	72.2	82.1
Mask-RCNN[6]	ResNet50+FPN	63.1	87.3	68.7	57.8	71.4
bottom-up						
Openpose[9]	-	61.8	84.9	67.5	57.1	68.2
HigherHRNet[11]	HRNet-W48	68.4	88.2	75.1	64.4	74.2
SWAHR[12]	HRNet-W48	72	90.7	78.8	67.8	77.7
CenterGroup[17]	HRNet-W48	69.6	89.7	76	64.9	76.3
Vector Field Regression						
DEKR[14]	HRNet-W48	70.0	89.4	77.3	65.7	76.9
ViTPose*[5]	ViTAE-G	80.9	94.8	88.1	77.5	85.9

Table 2. Performance Comparison on CrowdPose Dataset.

(PDR refers to the performance degradation rate of AP indicators on CrowdPose relative to AP on COCO)

method	PDR	AP	AP50	AP75	APE	APM	APH
Mask-RCNN[6]	0.0935	57.2	83.5	60.3	69.4	57.9	45.8
HigherHRNet[11]	0.0365	65.9	86.4	70.6	73.3	66.5	57.9
SWAHR[12]	0.0056	71.6	88.5	77.6	78.9	72.4	63

DEKR[14]	0.0386	67.3	86.4	72.2	74.6	68.1	58.7

3.2 Performance Analysis

In the top-down approach based on heat maps, the performance of HRNet[4] based on the separation model is significantly better than that of MaskRCNN[6] based on the joint model. The main reason for this phenomenon is that the detector and SPPE module in the separation model only need to complete their separate tasks, which is simpler than the joint task. And in the improvement of the training process and strategy, it is more convenient to optimize the intermediate process, making the output of the detector more accurate. From the table, it can also be seen that the separation model has significantly better estimation accuracy for small and medium-sized targets than the joint model. Through analysis, it can be concluded that firstly, due to the specificity of network tasks, the separation model uses more advanced backbone networks, and HRNet [4] has better feature extraction performance, especially multi-scale features, compared to ResNet; Secondly, the detection of all pedestrians in the separation model is unified into one size, which results in better keypoint prediction performance for small and medium-sized instances.

In the bottom-up approach, the algorithm groups and clusters all the predicted keypoints of the human body by the model, and then integrates them into different human bodies. The clustering method is the key to determining the performance of such methods. As shown in Table 1, clustering through correlated features generally performs well. This clustering method is based on extracting global features of key points for matching. Unlike integer linear programming and correlation vector fields, it can better preserve contextual information. At the same time, in SWAHR, different scales of key points can be located and features extracted by adjusting the standard deviation of the Gaussian kernel, which performs better in multi-scale scenes [12].

Vector field regression is more suitable for real-time prediction compared to heat maps. Firstly, due to its relatively simple model task, DEKR [14] can obtain keypoints through direct regression, while HigherHRNet [11] requires obtaining the probability of keypoints for the entire graph. Secondly, the vector field does not necessitate intricate post-processing procedures like non-maximum suppression, clustering correlation, and other computationally intensive processes. Meanwhile, vector field regression, due to its simple task, makes model optimization more convenient.

The accuracy degradation of models based on different methods also varies in situations with dense pedestrian traffic. As shown in Table 2, the performance loss of vector field regression and bottom-up methods based on feature association is relatively low in situations with a lot of pedestrian occlusion, especially the SWAHR performance degradation can be almost ignored [12]. This situation indicates that the way of associating features can improve clustering accuracy and instance segmentation by flexibly extracting ranges to combine contextual semantics. The top-down approach Mask-RCNN performs poorly in dense scenarios [6]. This is due to the fact that in the top-down approach, the model extracts key points from the bbox obtained by the detector. This method can affect the generation of the bbox in situations where pedestrian limb occlusion is large, causing the bbox to miss boxes and multiple boxes, resulting in a significant decrease in performance.

In recent years, Transformers have been widely applied in computer vision and have made remarkable progress, as shown in Table 1. Among them, the use of Transformer modules in Tokenpose [16] or the direct construction of backbone networks using ViT have been proven to achieve excellent results in multi person pose estimation tasks[5]. The self attention mechanism in Transformer and the idea of converting images into tokens can better address the multi-scale problem in human pose estimation. Moreover, ViTPose, which uses pure ViT to construct the skeleton, has also improved its computational speed during the inference process, reaching new heights in the balance of accuracy and efficiency, and has more potential in future multi person pose estimation [5].

4 Future Directions and Prospects

There are still many problems and directions to be studied in the current development process of multi person pose estimation. Future work can be carried out on the following aspects.

The scene with dense pedestrian occlusion and frequent occlusion increases the prediction difficulty for multi person pose estimation tasks. Pedestrian limbs often experience frequent occlusion and overlap in human intensive scenes. Current research has limited ability to deal with complex backgrounds, and algorithms with good performance in simple scenes often experience frequent false positives and omissions as pedestrian density increases. More algorithms specifically designed to handle crowded scenarios will be needed in future research.

Real time performance is a performance that must be improved in the implementation process of multi person pose estimation. Many networks nowadays sacrifice running speed while improving performance due to their complex feature fusion structures or complex algorithm logic. These networks often fail to meet real-time requirements when deployed on low computing power or embedded devices. Future research can focus more on improving operating speed while minimizing performance loss.

Network structures with stronger feature extraction capabilities, such as Transformer and Manba structures, can be combined with CNN or construct new networks to enhance the feature extraction ability of algorithms. Previously, many algorithms combined with the Transformer have achieved significant performance improvements, which further indicates that improving feature extraction capability is still the mainstream direction for improving algorithm performance. Future research can start by combining different network structures to solve problems from the perspectives of design and optimization.

5 Conclusions

In recent years, researchers have widely applied neural networks to multi-person pose estimation tasks. Based on the output type and data implementation logic of these networks, this research categorizes the methods into heatmap-based methods and vector

field regression-based methods. And heatmap-based methods can be further divided into top-down and bottom-up sub-methods. Methods based on deep learning have better performance than methods based on machine learning, but their performance still varies across different implementation logics.

The top-down method achieves the highest accuracy, but its performance is highly dependent on the detector's effectiveness. Among bottom-up methods, clustering by matching associated features is currently the most effective clustering strategy. The vector field regression method, offers high operational efficiency and is easy to optimize. In densely populated pedestrian areas, the top-down method exhibits the most significant performance degradation and is most sensitive to the level of pedestrian flow.

To clarify the development and characteristics of these different methods, this study summarizes typical approaches and analyzes their performance differences on the COCO2017 and CrowdPose datasets. This study not only compares the efficacy of heatmap-based and vector field regression-based methods but also conducts a comparative analysis of algorithms with different implementation logics. And it examines the varying degrees of performance degradation of different methods in dense scenarios and analyze the underlying reasons. Additionally, it addresses unresolved issues in MPPE, which include the performance degradation in dense scenarios and the demand for model lightweighting. This study aims to provide a reference for future algorithm improvements and hopes that future research can address the aforementioned issues.

References

1. Li, J., Zhang, S., Tian, Q., Wang, M., Gao, W.: Pose-guided representation learning for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. 44(2), 622–635 (2019)
2. Liu, Z., Wu, S., Jin, S., Liu, Q., Lu, S., Zimmermann, R., Cheng, L.: Towards natural and accurate future motion prediction of humans and animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10004–10012. IEEE, Piscataway (2019)
3. Fang, H. S., Xie, S., Tai, Y. W., Lu, C.: RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343. IEEE, Piscataway (2017)
4. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703. IEEE, Piscataway (2019)
5. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. Adv. Neural Inf. Process. Syst. 35, 38571–38584 (2022)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969. IEEE, Piscataway (2017)
7. Mao, W., Tian, Z., Wang, X., Shen, C.: FCPose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9034–9043. IEEE, Piscataway (2021)

8. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., Schiele, B.: DeepCut: Joint subset partition and labeling for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937. IEEE, Piscataway (2016)
9. Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., Sheikh, Y.: OpenPose: Real-time multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. 43(1), 172–186 (2019)
10. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. Adv. Neural Inf. Process. Syst. 30, 1–12 (2017)
11. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., Zhang, L.: HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395. IEEE, Piscataway (2020)
12. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13264–13273. IEEE, Piscataway (2021)
13. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578. IEEE, Piscataway (2019)
14. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14676–14686. IEEE, Piscataway (2021)
15. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6951–6960. IEEE, Piscataway (2019)
16. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S. T., Zhou, E.: TokenPose: Learning keypoint tokens for human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11313–11322. IEEE, Piscataway (2021)
17. Brasó, G., Kister, N., Leal-Taixé, L.: The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11853–11863. IEEE, Piscataway (2021)