2D Multi-Person Human Pose Estimation Based on Deep Learning

Ziyang Wang [1]

School of Tropical Agriculture and Forestry (School of Agriculture and Rural Affairs, School of Rural Revitalization), Hainan University, Danzhou, 571700, Hainan, China

wzyang2@outlook.com

Abstract. Human pose estimation leverages methods for computer vision to automatically find and recognize the principal joints of the human body. In the last few years, numerous deep learning techniques for estimating human posture have made great strides. Among these, since it forms the basis for 3D human pose estimation, 2D human pose estimation is crucial. Top-down and bottom-up approaches are the two broad categories into which 2D multi-person pose estimation methodologies can be separated. The former detects each object in the input data and then performs key point localization on each object individually. Its advantages include high accuracy and suitability for scenarios involving a single person or a small, dispersed group of people; however, its disadvantages include reliance on detection technology, high computational requirements, and poor real-time performance. The bottom-up method initially focuses on detecting all the key points contained in the input data, and then combines these key points based on their spatial relationships to form complete skeletons of different objects. The advantage is that it doesn't rely on object detection, making it suitable for scenarios with many people and occlusions; however, the disadvantage is that key point grouping is complex and prone to mis-matching. Furthermore, contrasting the two approaches' outcomes on the COCO dataset, this paper analyzes their performance in specific scenarios. It also discusses the remaining issues and future research directions in this field.

Keywords: Human pose estimation, Top-down approach, Bottom-up approach.

1. Introduction

Human pose estimation is a significant subject within the domain of computer vision that aims to automatically detect and localize the spatial location of important human body parts using information about human postures in pictures or movies. It has been used in a range of domains, including virtual reality, smart surveillance, and human–computer contact. Human pose estimation plays an important role as a basic task in the computer vision field.

Traditional methods of human pose recognition usually depend on manually extracted feature methods, which usually have the disadvantages of poor feature representation ability and poor ability to handle occlusion and pose variation (Yang & Ramanan, 2012). Its performance is greatly weakened in the face of a complex environment and crowd scene, and it is hard to meet the requirements of robustness and accuracy in the real scene.

With the development of computing ability and data amount, a deep learning based pose estimation method appeared. Compared with the traditional method, the deep learning based pose estimation method greatly improves efficiency and accuracy by omitting manual design of features (Pishchulin et al., 2012). After a long period of development, this method has become the mainstream method. 2D and 3D human posture estimation are the two categories into which human pose estimation may be divided according to the dimensions of the input data. (Liu et al., 2021). 2D human pose estimation directly estimates and discovers the two-dimensional coordinates of human body keypoints, and obtains the global posture. This work also helps 3D human pose estimation to determine the coordinates of key points in 3D space, so as to reconstruct the human body posture in the three-dimensional real world.

[1] https://orcid.org/0009-0008-6667-2295

2D human pose estimation provides reliable structural information for reconstructing three-dimensional poses. Therefore, high-quality 2D pose estimation is crucial for research related to human pose estimation. Depending on the detection method, the methods are often separated into two main categories: top-down and bottom-up.

Based on the aforementioned classification, this paper will introduce some representative models for both approaches, evaluate their performance on the COCO benchmark dataset, and discuss and explore future developments, aiming to provide researchers with new insights into this area.

2. Method

2.1. The Top-Down Approach

Top-down human pose estimation methods first use a human detector to locate each human instance in the image. Pose estimation is then performed on every human area that the detector has detected. On each detected region. Typically, object detection algorithms are employed to extract bounding boxes for each individual, followed by the application of a pose estimation network to predict keypoints for each cropped person image. Because it processes the human body region separately, the top-down approach often achieves higher accuracy in single-person situations or small-scale multi-person scenarios.

2.1.1 Masked Region-Convolutional Neural Networks (Mask R-CNN)

In early research, researchers used a two-stage approach: first, a pedestrian detector was used to detect pedestrians, and then a separate algorithm was used to estimate the pose of each individual pedestrian. This method has limitations, such as the detector sometimes only detecting partial body parts of pedestrians, resulting in lower accuracy. Furthermore, the design itself restricts detection efficiency and hinders further optimization. To address these issues, unified models like Mask R-CNN were developed.

This model builds upon the Faster R-CNN pedestrian detector through the addition of an additional network branch for single-person human pose estimation (He et al., 2017). This model first generates candidate regions (RoI), and subsequently employs a lightweight fully convolutional network (FCN) to generate segmentation masks for these regions, performing both classification and bounding box regression to identify the target category and precisely locate its position. This method reduces the number of parameters, enabling real-time pose estimation while improving localization accuracy. On the COCO dataset at the time, it achieved an instance segmentation accuracy of 37.1%, significantly outperforming methods such as FCIS, MNC, and SharpMask.

2.1.2 SimpleBaseline

As research progresses, the complexity of human pose estimation algorithms increases, leading to issues such as complex model structures and high computational resource consumption. Researchers aim to develop a simpler method that can achieve accurate pose estimation while ensuring higher feasibility and deployment efficiency.

SimpleBaseline was designed with this purpose in mind (Xiao, Wu, & Wei, 2018). This method first uses a standard convolutional neural network (ResNet) to extract feature maps. It directly adds deconvolutional layers on top of it to reconstruct the low-resolution feature map into a high-resolution heatmap. In the heatmap representation, individual channels correspond to distinct keypoints, and the peak value within each channel denotes the predicted location of the associated keypoint.

This method, despite its simple architecture, achieves a top score of 73.7% on the COCO dataset (using ResNet-152 and additional data), surpassing most complex multi-stage methods of the time, while also offering stable training and efficient inference.

2.2. The Bottom-Up Approach

In bottom-up approaches, the process begins with detecting all body keypoints, grouping these points through their geometric relationships in order to assemble the skeletal configuration of each individual. Because its key point detection is performed in a single step, without relying on first detecting individual body regions, bottom-up pose estimation methods exhibit more stable performance when processing high-density crowds, and generally have shorter processing times, resulting in better real-time performance. The following section will introduce some representative methods.

2.2.1 OpenPose

In light of the rapid progress in deep learning, by employing ResNet, researchers hope to enhance the real-time performance and resilience of human keypoint localization. To achieve an end-to-end, multi-person, high-accuracy, and real-time human pose estimation method without requiring prior detection of individual bounding boxes, researchers proposed the OpenPose method (Cao et al., 2017).

This method proposes a dual-branch neural network architecture based on nonparametric representation, which predicts both the heatmap of human keypoint locations and the Part Affinity Fields (PAFs) separately. The system uses an analytical algorithm to perform local extremum detection on the predicted heatmap to determine the positions of key points. Then, it calculates the correlation coefficient for each pair of key points by integrating along the line connecting them using PAF, and uses this correlation coefficient for clustering, thus combining these key points to form a complete human body skeleton structure. The introduction of the "Part Affinity Field" (PAF) concept laid a crucial foundation for subsequent bottom-up human pose estimation research. This method achieved an accuracy of approximately 61.8% on the COCO 2016 validation dataset, demonstrating comparable or even slightly better accuracy than existing multi-stage network approaches, while also offering faster inference speed and enabling real-time pose estimation.

2.2.2 PifPaf

The accuracy of Keypoint detection and linking in OpenPose is quite poor in complex scenarios, and thus, it mismatches a lot. Particularly, when the input is low resolution and the scenario is complex (i.e., there is more than one person in the scenario) and there is a high amount of occlusion, researchers have proposed the PifPaf method(Kreiss, Bertoni, & Alahi, 2019).

PifPaf is very similar to OpenPose in its overall architecture. OpenPose uses PAF to record the connections between the main points, and vector fields describe the connection direction. While PifPaf uses two head networks, and since OpenPose researchers add on top of OpenPose, more accurate Part Intensity Fields (PIFs) to localize the keypoints, and also use PAFs to cluster the keypoints.

This improvement enables PifPaf to provide data processing performance close to conventional methods at high resolution and vastly better performance at low resolution. The improvement yields better results on the COCO 2017 val set and on the CrowdPose dataset. Specifically: 66.7% on COCO with a ResNet-101 backbone, and substantially better AP than current cutting-edge techniques. In congested settings, use OpenPose and Mask R-CNN.

3. Results comparison

One well-liked benchmark dataset for computer vision multi-person posture estimation is the COCO (Common Objects in Context) dataset.

Coco focuses on multiple objects in the complex scene. The COCO annotation focuses on the presence of multiple objects in images. That is, a single image may contain many instances of objects and a complicated background. Models should address the challenges of multiple objects, occlusion, varied scales in the same model, etc. The dataset includes 80 common object classes, with more than 200k images and 250k instance annotations, with rich scene context and

high-quality manual annotation (Kong, Qin, & Zhang, 2023). The COCO is a benchmark platform for Object detection, segmentation, keypoint, image understanding, etc. Average precision (AP) is the standard assessment statistic for the COCO dataset. AP is calculated with Object Keypoint Similarity (OKS).

OKS is a measure of how much the predicted keypoints must be allowed to deviate with respect to the ground truth. It is aimed to reflect the following describes how the anticipated keypoints and the ground reality are comparable.

$$OKS_p = \frac{\sum_i \exp\left(-\frac{d_{pi}^2}{2S_p^2\sigma_i^2}\right)\delta(v_{pi}>0)}{\sum_i \delta(v_{pi}>0)} \quad (1)$$

Here, d_{pi} symbolizes the Euclidean distance at the i-th key point position between the actual point and the expected point. s_p shows how big the target item is, σ_i is a constant coefficient for several important points, and v_{pi} is the i-th key point's visibility indication. The function $\delta(\ )$ is an indicator function, which only takes the value 1 when a key point is visible.

Based on OKS, the average precision (AP) can be further calculated. In single-person scenarios, and in top-down human pose estimation, the definition of AP is as follows:

$$AP=\frac{\sum_p \delta(OKS_p>T)}{\sum_p 1} \quad (2)$$

In this formula, $\delta(OKS_p > T)$ is considered a correct detection if the OKS value of the p-th prediction is greater than the set threshold T. In this formulation, the numerator indicates the count of accurate predictions, whereas the denominator represents the overall number of predictions made.

In bottom-up human pose estimation, the definition of AP is as follows:

$$AP = \frac{\sum_m \sum_p \delta(OKS_p>T)}{\sum_m \sum_p 1} \quad (3)$$

Here, m represents different human instances, and the calculation is performed by statistically analyzing all predicted instances. The final COCO key point detection metric is obtained by averaging the results across different OKS thresholds.

In addition to AP, the evaluation metrics for each model also include AP values at different Intersection over Union (IoU) thresholds, such as AP_{50} and AP_{75}, representing the AP values when IoU is 0.5 and 0.75, respectively. They also include AP values for different object size categories, such as AP_M (medium-sized objects) and AP_L (large-sized objects), which represent the detection performance for different object sizes.

In addition to the model previously mentioned, this research also compares a number of other popular methods on the COCO dataset in order to fully assess the effectiveness of other approaches. The findings are shown in Table

Table 1: Performance comparison on the COCO dataset.(Li, Wang, & Zhang, 2024)

method	AP	AP_{50}	AP_{75}	AP_M	AP_L
Top-down approach					
Mask R-CNN	63.1	87.3	68.7	57.8	71.4
SimpleBaseline	73.7	91.9	81.8	70.3	80.0
CPN	72.1	91.4	80.0	68.7	77.2
Bottom-up approach					
OpenPose	61.8	84.9	67.5	57.1	68.2
PifPaf	66.7			62.4	72.3
CenterGroup	69.6	89.7	76.0	64.9	76.3

As illustrated in Table 1, the top-down human pose estimation achieves better performance on the COCO dataset. This is likely because the top-down approach provides strong prior information through human body detection, allowing key point prediction to be performed in relatively clean, localized areas, thus effectively reducing background interference and confusion between multiple human bodies. Compared to bottom-up methods that simultaneously predict and associate all key points across the entire image, top-down methods exhibit better convergence. Furthermore, given the high proportion of small objects in the COCO dataset, top-down methods can preserve more detailed information during key point prediction by cropping and scaling the bounding boxes. This two-stage processing approach is better suited to address challenges such as the wide range of object sizes, significant object overlap, and complex backgrounds found in the COCO dataset.

4. Future outlook and direction

Although human posture estimation is acknowledged as an important area of study in computer vision, there are still some issues that need to be resolved.

4.1. Robustness Issues in Diverse Environments

While current mainstream human pose estimation models perform well on standard datasets, their accuracy significantly decreases in real-world scenarios involving strong or low light conditions, occlusions, cluttered backgrounds, or varying camera angles. This sensitivity to environmental conditions limits the practical application of these models in areas such as security monitoring, autonomous driving, and AR/VR. Researchers are exploring ways to integrate data from multiple sources, such as RGB images, depth information, thermal imaging, and inertial sensors, in order to enhance the models' accuracy across diverse domains and varying data distributions, thus enabling their broader application. For example, Dayarathna et al. conducted experiments on fusing LWIR image features with visible light images (Dayarathna et al., 2023). Their study facilitated the generation of synthetic visible-light images from LWIR data, thereby enabling accurate even in the absence of real visible-light images, and consequently enhancing the method's applicability and robustness.

4.2. Ability to Handle Multi-Player Interactions and Occlusion Issues

In multi-person pose estimation, the accuracy significantly decreases when there is occlusion, overlapping, or complex interaction between individuals. This is especially true in crowded scenes, where the model has difficulty accurately distinguishing and identifying the body positions and movements. Existing top-down and bottom-up approaches both have certain limitations (Qi et al., 2024). Current research mainly focuses on incorporating strategies such as spatial-temporal modeling and attention mechanisms. The former approach utilizes temporal information to track individual motion trajectories, thus maintaining the continuity of pose estimation even when occlusion occurs; the latter employs an attention mechanism to dynamically focus on visible key point regions, thereby improving recognition accuracy. Both research directions contribute to enhancing the system's adaptability in complex interactive scenarios.

4.3. Balancing Efficiency and Practicality

Many high-precision models currently rely on large neural networks and high-performance computing resources, making them difficult to deploy on resource-constrained devices, thus limiting their application in edge computing and mobile devices. Therefore, improving the computational efficiency and reducing the memory usage of these models while maintaining accuracy has become an urgent issue that needs to be addressed. Future research will focus on developing lightweight pose estimation networks that can run efficiently on mobile and embedded platforms. Tan et al., for instance, suggested a platform-aware neural architecture search technique (Tan et al., 2019). This method incorporates the actual inference latency on mobile devices into the

optimization objective, attaining an equilibrium between precision and effectiveness. By employing a hierarchical, factorized search space and a reinforcement learning controller, the resulting network achieves 75.2% accuracy on ImageNet and has a latency of only 78ms on a Google Pixel phone, outperforming both MobileNetV2 and NASNet-A in both speed and accuracy. Furthermore, it demonstrates excellent performance on COCO object detection tasks, providing valuable insights for the design of lightweight and efficient neural networks.

5. Conclusion

Human posture estimation is an important field of study in contemporary computer vision research. This article mainly discusses 2D multi-person human pose estimation techniques, classifying them into top-down and bottom-up. The former first recognizes the people in the image, generating bounding boxes for each person, and then performs individual pose estimation. This method processes each person independently, achieving high accuracy in keypoint localization, and performs well in scenarios where people are spaced far apart. However, it is inefficient in large-scale scenes or scenarios with significant occlusion between people. The latter method initially detects the keypoints of all individuals present in the image, matches and combines these keypoints to identify individual pedestrians. This method is fast and suitable for real-time human pose estimation, and has better robustness in crowded scenes; however, it suffers from issues such as high computational complexity and potential matching errors. This paper also introduces and analyzes representative models from both types of methods. Subsequently, it compares the performance of several commonly used models, including the introduced models, on the COCO dataset, and finds that top-down human pose estimation methods achieve better performance. Based on these results, this paper analyzes some possible reasons. It concludes by outlining some of the problems that still plague human pose estimate research and talking about potential avenues for further study.

References

[1] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7291–7299).

[2] Dayarathna, T., Muthukumarana, T., Rathnayaka, Y., Denman, S., De Silva, C., Pemasiri, A., & Ahmedt-Aristizabal, D. (2023). Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study. Expert Systems with Applications, 213, 119139.

[3] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961–2969).

[4] Kong, Y., Qin, Y., & Zhang, K. (2023). A review of deep learning 2D human pose estimation methods. Journal of Image and Graphics, 28(07), 1965–1989.

[5] Kreiss, S., Bertoni, L., & Alahi, A. (2019). Pifpaf: Composite fields for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11977–11986).

[6] Li, J., Wang, D., & Zhang, S. (2024). 2D human pose estimation based on deep learning: Current status and prospects. Chinese Journal of Computers, 47(01), 231–250.

[7] Liu, S., He, N., Wang, C., & Yu, H. (2021). Research progress of 2D human pose estimation. In (eds.) Proceedings of the 25th Annual Conference on New Network Technologies and Applications, Network Application Branch, China Computer Users Association (pp. 255–258). Beijing Key Laboratory of Information Service Engineering, Beijing Union University.

[8] Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., & Schiele, B. (2012, June). Articulated people detection and pose estimation: Reshaping the future. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3178–3185). IEEE.

[9] Qi, X., Xu, C., Zeng, F., & Yao, D. (2024). A review of human pose estimation based on deep learning. In (eds.) Proceedings of the 28th Annual Conference on New Network Technologies and Applications, Network Application Branch, China Computer Users Association (pp. 280–283). Beijing Key Laboratory of Information Service Engineering, Beijing Union University; Beijing Laboratory of Brain and Cognitive Intelligence, School of Robotics, Beijing Union University.

[10] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). MnasNet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2820–2828).

[11] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 466–481).

[12] Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2878–2890.