A Review on Human Pose Estimation Based on Deep Learning

Wenhao Chen

School of Mathematics and Information Science, South China Agricultural University, Guangdong, China

hongzhen@ldy.edu.rs

Abstract. With the rapid development of computer vision technology, deep learning-based human pose estimation has become a hot topic of research. This review outlines the progress made in this field in recent years, with a particular focus on the development of single-person and multi-person pose estimation. Single-person pose estimation primarily focuses on identifying and locating the joints of an individual, while multi-person pose estimation further extends to simultaneously recognizing the poses of multiple individuals. The article begins by introducing the basic concepts of pose estimation, then discusses in detail the application of deep learning models in single-person and multi-person pose estimation, as well as the advantages and disadvantages of the existing modules. In addition, the limitations of current models are analyzed at the end of the paper, and possible future optimization directions are explored. The aim of this article is to provide researchers with a comprehensive perspective to understand current technological trends and potential innovation points.

Keywords: Computer Vision, Single-Person Pose Estimation, Multi-Person Pose Estimation, Deep Learning.

1. Introduction

Human pose estimation refers to the task of identifying and locating the positions of human joints from images or videos, which is an important task in the field of computer vision and is significant for understanding human behavior, interaction, and activities. With the emergence and continuous popularity of deep learning technology, the human pose estimation methods based on deep learning have become the forefront of research and the need for development. These methods utilize the powerful feature extraction capabilities and self-learning abilities of deep neural networks, significantly improving the robustness and accuracy of pose estimation and further advancing the development of human pose estimation.

In existing research on single-person pose estimation, researchers have developed various effective deep learning architectures, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), which can precisely predict the joint positions of an individual. However, these methods often rely on a lot of annotated data to train the models and use complex network structures to capture changes in human posture.

When multiple individuals are present in a scene, the complexity of human pose estimation increases significantly. Multi-body pose estimation must not only recognize the joints of each person but also distinguish the joints belonging to different individuals. This requires algorithms to handle issues such as occlusion, overlap, and changes in viewpoint. In recent years, researchers have proposed various strategies to address these challenges, including using more complex network structures, improved data augmentation techniques, and innovative loss functions.

This article reviews the research progress in deep learning-based body pose estimation, focusing on the development of single-people and multi-people pose estimation and the existing modules. The paper first reviews the basic concepts and challenges of pose estimation, then discuss in detail the application of deep learning in this research area. Next, the paper analyzes the limitations of current methods and propose possible future research directions. The paper hopes to provide researchers with a clear technological roadmap to guide future research and development work.

2. Analysis of Single-person Pose Estimation

In the field of computer vision, a key task is single-person pose estimation, which aims to detect and recognize specific body parts of a person from images or videos, such as the head, torso, arms, and legs, etc. These parts of the body are usually defined as keypoints, and the key to the task is to understand and interpret human movement and structure through these keypoints. This task is of great significance to fields such as human-computer interaction, gait analysis, video surveillance, entertainment games, and health monitoring. Currently, methods for single-human pose estimation can be mainly divided into two categories: traditional methods and deep learning-based methods.

Traditional Methods: These methods are usually based on graph structures and deformable part models, designing 2D human part detectors, using graph models to establish connectivity between parts, and combining constraints of human kinematics to optimize the graph structure model for body pose estimation. However, the disadvantage is that they rely on manually set features, such as methods based on feature descriptors, HOG and SIFT features[1], which cannot fully utilize image information, and when the human pose changes significantly, the part model struggles to accurately depict and express this deformation.

Deep Learning-Based Methods: With the development of deep learning and the advent of the big data era, the field of computer vision has successfully adopted the concept of deep learning and has seen rapid development. Deep learning-based body pose estimation methods mainly use CNN to automatically extract human pose features from images and perform pose estimation[2]. Compared to the traditional methods, CNN can obtain features that are richer in semantic information and can obtain the full context of each feature under different receptive fields, thus eliminating the dependence on the design of part model structures.

In deep learning-based methods, single-people pose estimation algorithms can be further divided into methods based on linear regression and methods based on heatmap detection. Linear regression methods directly map from body features to body parts, training the network to directly predict the coordinates of each joint, while heatmap detection methods predict the approximate positions of body parts and joints, which are supervised by heatmaps representing these parts and joints.

Overall, research in single-person pose estimation is moving towards more accurate and real-time development, and deep learning methods, especially models based on CNN and Transformer, have shown great potential and advantages in this field [3]. The following will discuss the development of methods based on linear regression and heatmap detection, as well as the advantages and disadvantages of specific modules.

2.1. Based on Linear Regression methods

Linear regression-based single-person pose estimation methods were a mainstream technical means before the emergence of deep learning technology. These methods typically regard pose estimation as a regression problem, that is, the direct mapping from image features to keypoint positions. Early representative work includes the DeepPose model proposed by Toshev et al. [4], which was one of the first algorithms to use deep learning to human pose estimation. It uses a convolutional neural network (CNN) to extract features, employs AlexNet as the base network structure, and directly regresses the coordinates of the joint points through a designed cascade network. It can handle complex human pose changes, but due to its regression-based approach, its detection effect on occluded and overlapping joint points is limited.

With the development and practical application of deep learning technology, traditional linear regression methods have gradually been replaced by deep learning-based pose estimation methods. Deep learning methods can automatically learn high-level features of images, providing more accurate and robust pose estimation. For example, the Convolutional Pose Machine (CPM) model [5] predicts the position of keypoints through a multi-stage network structure, gradually refining the prediction.

Nevertheless, linear regression methods still have their value in certain situations, especially in applications with small amounts of data or where fast inference is required. Due to their simplicity and efficiency, linear regression models can serve as a baseline model or be combined with other deep learning methods to improve the accuracy and efficiency of pose estimation.

2.2. Based on Heatmap Detection methods

Single-person pose estimation methods based on heatmap detection have seen rapid development in recent years, mainly thanks to the progress of deep learning technology. These methods typically represent the position of keypoints as heatmaps and determine the position of keypoints by predicting heatmaps. Here is an introduction to some representative modules:

Hourglass Network[6]: The Hourglass network is a network structure for single-person pose estimation. It adopts an hourglass-shaped network design, effectively capturing the spatial information of human body joints through multi-scale feature fusion and repeated downsampling and upsampling processes.

AlphaPose[7]: AlphaPose is an outstanding human pose estimation framework, the core of which is the Region Multi-Person Pose Estimation (RMPE) model. This framework performs excellently in both single and multi body pose estimation tasks.

HRNet[8]: High-Resolution Network (HRNet) maintains high-resolution feature map information by parallelizing multi-resolution sub-networks and enhances high-resolution features through multi-scale fusion, thereby improving the performance of pose estimation.

UniPose[9]: UniPose is a unified human pose estimation framework that combines the cascade method of atrous convolution and the Atrous Spatial Pyramid Module to determine keypoint positions and human bounding boxes, achieving high-precision estimation.

2.3. Summary

Linear regression-based single-person pose estimation methods have made significant progress under the impetus of deep learning technology, but they are still in the process of continuous development and optimization; heatmap detection-based methods have demonstrated better robustness and accuracy in handling complex scenes, occlusions, and pose variations. Future research may focus on how to combine the advantages of traditional methods and deep learning approaches, as well as how to deal with more complex scenarios and pose changes. Table 1 compares the strengths and weaknesses of specific modules for single-person pose estimation.

Table 1. Comparison of Specific Modules for Single-Person Pose Estimation

Module	Advantages	Disadvantages
DeepPose	For the first time, deep learning was introduced into human pose estimation, which can handle pose variations and occlusions to a certain extent.	Facing complex scenes, efficiency decreases.
Hourglass Network	It can effectively handle occlusion issues and predict the spatial relationships of keypoints.	It requires a multi-stage network structure, which is computationally expensive.

Table 1. (continued).

AlphaPose	It has achieved excellent performance on multiple datasets, especially in dealing with occlusions and complex scenes.	As the number of people in the images increases, the computational load increases and the speed becomes slower.
HRNet	It maintains high-resolution feature maps, thus avoiding the information loss caused by low-resolution feature maps. It performs parallel computing at multiple resolutions, improving computational efficiency, and can adapt to different tasks by adding branches.	It requires substantial computational resources for training and inference, has a large number of parameters, and needs more data and computational resources for training. It is also sensitive to the size of the input images.
UniPose	It is the first to achieve the task of handling text and visual prompts within a single model.	It has a wide range of applications but may require further optimization to improve performance in specific scenarios.

3. Analysis of Multi-person Pose Estimation

Compared to the single-person pose estimation, the multi-person pose estimation is more complex and faces more challenges, as it involves detecting and locating the body keypoints of multiple individuals in images or videos, while dealing with occlusions, crossings, and varying human body shapes. In recent years, deep learning-based multi-person pose estimation methods have made significant progress, mainly divided into top-down and bottom-up approaches.

Top-down approach: This method detects and recognizes all individuals in the image at first, then uses a human detector to distinguish each person, generates a bounding box for each individual, and performs single-person pose estimation within the bounding box. This approach relies on accurate human detection, and the accuracy of the human detection will directly affect the effectiveness of subsequent pose estimation.

Bottom-up approach: This method first detects all keypoints in the image, then uses some form of clustering or graph optimization to assign keypoints to different human bodies, thereby determining which keypoints belong to each individual. This approach does not depend on initial human body detection, making it more robust in handling occlusions and intersections.

As research progresses, future multi-person pose estimation methods may focus more on improving accuracy, robustness, and real-time performance, while also exploring how to apply these technologies to a broader range of applications, such as action recognition, virtual reality, and human-computer interaction. The following will discuss the development of the two approaches and the advantages and disadvantages of specific modules.

3.1. Top-down methods

Top-down multi-person pose estimation algorithms have seen significant development in recent years, and here are some representative algorithms along with their advantages and disadvantages:

Faster R-CNN + CPM[10]: Papandreou et al. proposed a method in their paper that combines Faster R-CNN for human detection and CPM (Convolutional Pose Machine) for pose estimation. This module integrates the efficient object detection of Faster R-CNN with the pose estimation capabilities of CPM. The method first utilizes Faster R-CNN to detect humans, segmenting each individual body, and then employs CPM for keypoint estimation of pose. The advantage of this approach lies in its ability to handle complex scenes and is widely used in scenarios that require both human detection and pose estimation. However, the drawback of Faster R-CNN + CPM is its high dependency on the accuracy of the detection phase, and the computational cost increases with the number of people detected.

Region Multi-Person Pose Estimation (RMPE)[11]: Fang et al. proposed a regional multi-person pose estimation framework, which includes three important components. The first is the Symmetric Spatial Transformer Network, it can extract high-quality individual regions from inaccurate human detection boxes, thereby promoting the accuracy of pose estimation. The second is Parametric Pose Non-Maximum Suppression , a technique that can be used to eliminate redundant pose estimations. The third is the Pose-Guided Proposals Generator, which can produce a large number of training data samples, thereby enhancing the training dataset.

Structure-Preserving Pose Estimation (SPLP)[12]: The SPLP model proposed by Pishchulin et al. can simultaneously detect the number of individuals and their poses in an image and has good adaptability to occluded scenes. The advantage of this method is its effective handling of occlusion issues, but the downside is its high algorithmic complexity and longer computation time.

3.2. Bottom-up methods

Bottom-up multi-person pose estimation algorithms primarily detect all keypoints in the image and then use clustering algorithms to assign these keypoints to different individuals, thereby solving the multi-person pose estimation problem. Here are some bottom-up multi-person pose estimation algorithms and their advantages and disadvantages:

Openpose[13]: Openpose is a popular bottom-up multi-person pose estimation system that uses Part Affinity Fields (PAFs) to predict connections between keypoints and then assigns keypoints to different individuals through a graph-cut algorithm. The advantage of Openpose is its ability to handle multi-person occlusions and complex scenes, but its disadvantage is the high computational load, especially when processing high-resolution images.

Openpifpaf[14]: Openpifpaf is a bottom-up multi-body pose estimation method based on composite fields that performs well in urban traffic scenes. The advantage of Openpifpaf is its adaptability to low-resolution and crowded scenes, but the disadvantage is the potential need for more computational resources.

DeeperCut[15]: DeeperCut is a bottom-up approach using deep learning that improves the accuracy of pose estimation through an improved convolutional neural network and graph-cut algorithm. The advantage of DeeperCut is its ability to provide precise pose estimation, but the high computational cost is the disadvantage.

3.3. Summary

The methods of Top-down multi-person pose estimation still face challenges in dealing with occlusions, complex scenes, and computational costs. Future research may focus on improving the robustness of algorithms, reducing the demand for computational resources, and enhancing adaptability to complex scenes. Bottom-up algorithms are generally more robust in handling occlusions and complex scenes, but their disadvantage lies in the high computational cost, especially when dealing with a large number of keypoints and complex clustering issues. Future research may focus on how to improve the efficiency and accuracy of these algorithms and how to better adapt to different application scenarios. Table 2 compares the advantages and disadvantages of specific modules for multi-person pose estimation methods.

Table 2. Comparison of Specific Modules for Multi-Person Pose Estimation

Module	Advantages	Disadvantages
Faster R-CNN + CPM	Capable of handling complex outdoor scenes; there are many points in the algorithm framework that can be optimized, providing a broad space for optimization.	High computational load, slower speed when dealing with small or dense targets.

Table 2. (continued).

Region Multi-Person Pose Estimation (RMPE)	Able to handle inaccurate human bounding boxes.	The algorithm framework is relatively complex, which may require more tuning and computational resources.
SPLP (Structure-Preserving Pose Estimation)	Maintains the consistency of the body structure during the estimation process, which helps to improve the accuracy of human pose estimation.	May require additional costs, and fine-tuning of parameters is needed to achieve optimal performance.
OpenPose	Real-time multi-person 2D pose estimation, suitable for human-computer interaction.	High memory consumption, requiring higher hardware configurations; detection performance may not be good in special scenarios.
Openpifpaf	A method based on composite fields, suitable for urban mobility scenarios.	May require more optimization for handling dense crowds or complex backgrounds.
DeeperCut	Graph optimization strategies improve performance and speed.	High computational complexity, which may require more computational resources.

4. Challenges and Difficulties

Deep learning-based body pose estimation is a popular and significant research direction in the field of computer vision. Despite the remarkable progress made in recent years and the emergence of many efficient and lightweight modules, there are still some challenges and difficulties. Here are some of the current hot topics and urgent challenges that need to be addressed:

1） Occlusion problem: In multi-person pose estimation, when body parts are occluded by each other, it is difficult to distinguish the keypoints belonging to each individual.In such complex scenarios, the algorithm is required to accurately identify the ownership of each keypoint and the posture of the occluded parts.

2） Fine-grained and accuracy: Precise estimation of detailed parts such as hands and faces is a severe challenge. The variations of detailed parts are more diverse and complex than the overall posture changes of a person, and these detailed parts are more prone to occlusion.

3） Cross-modal fusion: How to effectively combine data from different sensors, such as optical sensors, radar, infrared sensors, etc., to improve the accuracy and robustness of human pose estimation is also a hot research topic.

5. Conclusion

Human pose estimation is a hot research in computer vision. This article reviews the research progress and representative modules in deep learning-based human pose estimation. By introducing single-person pose estimation and multi-person pose estimation, as well as providing an overview and comparison of various modules, it systematically elaborates on the current state of deep learning-based body pose estimation and the challenges it faces. Future research directions may focus on the lightweight design of module algorithms to enhance real-time performance; the transformation from two-dimensional to three-dimensional, providing richer application scenarios for virtual reality, augmented reality, and animation production; and unsupervised and semi-supervised learning to reduce reliance on large amounts of annotated data, utilizing unannotated data for training to improve the model's generalization capabilities.

References

[1] Wang, K., et al. (2024). A Review of 2D Human Pose Estimation Based on Deep Learning. Journal of Zhengzhou University (Science Edition) 56.04: 11-20.

[2] Zhang, G.P., et al. (2022). Research Progress of Deep Learning Methods in 2D Human Pose Estimation. Computer Science 49.12: 219-228.

[3] Shi, X. R., (2023). Research on 3D Human Pose Estimation Based on Deep Learning. East China Normal University, MA thesis.

[4] Yu, B. L., et al. (2020). A Multi-Target Human Skeleton Node Detection Algorithm Based on DeepPose and Faster RCNN. Journal of the Chinese Academy of Sciences 37.06: 828-834.

[5] Wang, J.Y., et al. (2022). A Method for Identifying Divers' Operating Postures Based on Convolutional Pose Machine. Journal of Zhejiang University (Engineering Science) 56.0: 26-35+46.

[6] Xiao, Y.B. (2023). Research on Human Pose Estimation, Behavior Analysis, and Application Based on Deep Learning. Beijing University of Posts and Telecommunications, PhD dissertation.

[7] Jia, Y.G., et al. (2023). Standardized Evaluation of Running Action Based on Alphapose. Computer Era. 08: 117-120.

[8] Liao, J.H., et al. (2024). Lightweight HRNet: A Ligtweight Network for Bottom-Up Human Pose Estimation. Engineering Letters 32. 3.

[9] Artacho, B. and Savakis A. (2020). UniPose: Unified human pose estimation in single images and videos. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[10] Chen, G.L., and Pang Y.S. (2024). An Improved Faster RCNN Algorithm for Micro-Operation Spatial Target Detection. Sensors and Microsystems 43.03: 144-147+151.

[11] Li, J. (2021). Research on Bottom-Up Multi-Person Pose Estimation Methods. University of Science and Technology of China, PhD dissertation.

[12] Sun, L. (2022). SPLP: A Certifiably Globally Optimal Solution to the Relative Pose Estimation Problem Using Points and Line Pairs." Journal of Interconnection Networks 22. Supp02.

[13] Li, X.T. (2023). Research on Human Action Recognition Based on Deep Learning. Shanxi University, MA thesis.

[14] Micheal D., et al. (2024). An Interpretable Modular Deep Learning Framework for Video-Based Fall Detection. Applied Sciences 14.11.

[15] Eldar I., et al. (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. CoRR abs/1605.03170.