# Research on Multi-Person Pose Estimation Technology

Hongyan Wang*
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: whyanon@163.com
*corresponding author

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

***Abstract*—Human pose estimation is a hot topic of computer research in recent years, which promotes the progress of society and brings many conveniences to people's lives. Fom traditional methods to the mainstream deep learning-based methods, the primary approach in deep learning involves the use of convolutional neural networks to reduce computational complexity and improve network accuracy, but because the network structure is too deep to improve the accuracy, the trained model parameters are also very large, and it is very dependent on the input of hardware equipment. At this time, the lightweight human pose estimation can solve this problem very well. This paper mainly describes the knowledge of convolutional neural network in detail and compares it with traditional image algorithms. The OpenPose model is a classic model based on convolutional neural network that can well achieve single-person and multi-person human pose estimation model, but because the convolution kernel in its network structure is too large to increase the amount of calculation, this paper proposes three improvements to the network structure of the conventional OpenPose model. Finally, the precision of the model is improved by about 40%, which verifies the feasibility of lightweight human body posture estimation research.**

***Keywords-Convolutional Neural Network; Network Structure; OpenPose Model; Multi-person Human Pose Estimation***

## I. INTRODUCTION

In recent years, computer vision related technologies have developed rapidly. Human posture estimation [1] involves the most basic method of human posture research, and has become a hot spot for many scholars. Its primary objective is to identify the spatial coordinates of human joints and key body parts [2][3] within an image, enabling the extraction of partial or comprehensive limb information. This information is crucial for assessing and interpreting human body movements and behaviors. Due to many factors such as shooting angle, lighting, and environment, it is difficult for traditional human pose estimation algorithms to achieve satisfactory results. Given the extensive exploration of deep learning, particularly convolutional neural networks the utilization of such networks in human pose estimation has progressed swiftly, the traditional manual extraction of features is replaced by the method of learning features by convolutional neural networks, thereby realizing end-to-end optimization [4], although the current method based on convolutional neural networks has emerged as the mainstream method for human pose estimation, but there are still some core problems that have not been solved. For example, the existing research work mainly focuses on enhancing the accuracy of human pose estimation methods, resulting in progressively intricate network models. However, this emphasis on accuracy sometimes neglects the crucial trade-off between precision and processing speed in the network.

## II. RELATED WORKS

### *A. Survey of Pose Estimation*

There are two ways to estimate two-dimensional pose, one is to measure all the heads, left and right hands, knees, etc. from the bottom up [5][6], and then connect all the joints with the human body and then combine them together. The second is from the top down, transforming the pose estimation of multiple people into the pose estimation of multiple individuals, typical of CPM, Hourglass, CPN, Simple Baselines, HRNet,

DOI: 10.2478/ijanmc-2023-0075

MSPN, etc [7]. OpenPose is currently the most popular method of bottom-up multiplayer pose assessment [8]. First, the OpenPose network places multi-person photos on the previous level for feature extraction, and then inputs the features to two parallel branches, and first obtains a set of credibility maps from one branch, each confidence map represents the key points of the human body bone composition [7][8]. The second branch is used to predict the importance of other parts. All that remains is to refine the predictions of each branch, make a bipartite map of the credibility of each part, and then use the PAF value to stitch together the weaker parts of the dichotomy to obtain a rough outline of the human body and piece it together into a person.

With the advancement of computer equipment and human pose estimation technology, the utilization of conventional convolutional neural network algorithms in deep learning has been caused by complex convolutional convolution and large computational capacity to meet the requirements of the people and market demand under the current social form [9].

## B. *The OpenPose Mode*

Cao et al. [10] introduced the OpenPose model, employing a technique known as the "local affinity field". This method effectively encode the position and orientation of limbs so that the connections between key points can be appropriately combined. The motion classification model uses neural networks to deeply fuse and classify human body key points output by the human posture estimation model, thereby achieving real-time monitoring of human motion.

## C. *Currently*

At present, how to maintain the high performance of the model and make it achieve lighter mass has become a hot spot in current research. The OpenPose model extracts the feature map in the VGG network [11], a convolutional neural network is employed to analyze the reliability of the key points and the affinity vector field associated with these key points, and uses the Hungarian algorithm to match and optimize the key points [12][13]. Although the network complexity is reduced, the performance improvement is not significant, and the amount of operation is also increased, so in order to reduce the network complexity, reduce the network parameters, reduce the amount of calculation, on the basis of OpenPose, a lightweight model light OpenPose is proposed to realize the lightweight human pose algorithm [9][13], so that the human pose estimation in practical applications on the basis of liberating manpower to achieve better results. The components of lightweight human pose estimation are illustrated in Figure 1.

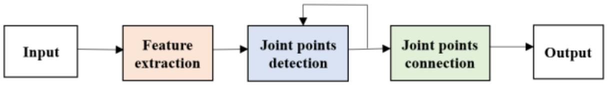


Figure 1. Components of lightweight human pose estimation

# III. METHODS AND MATERIAL

## A. *Human Skeletal Coding*

The distinctive attributes of different body parts can be condensed into either 18 or 25 key feature points. The human skeleton, formed by these feature points, effectively portrays the body's posture. The accuracy of human movement can be assessed based on the angles between the joints of the human body, and will not be affected by factors such as human body type, skin color, and clothing. This article is to use this feature to identify motion.

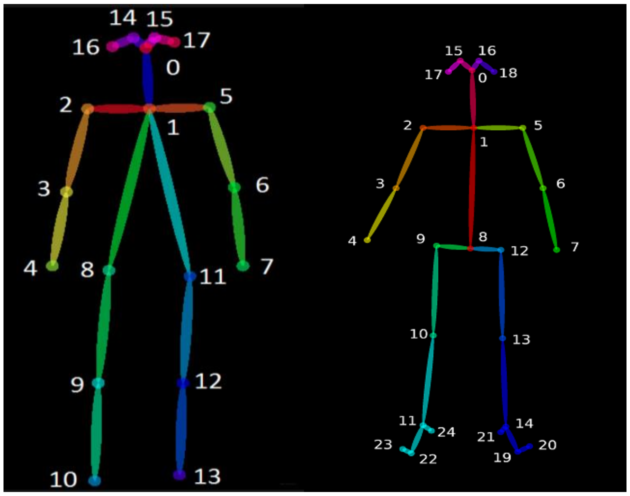


Figure 2. COCO (left) and OpenPose (right) coding diagrams

Deep learning serves as the foundation for conducting numerous descriptions of human body posture information. Currently, within the realm of human pose estimation, the labeling of data is also the marking of bone as the key point. The coding method in this article is mainly shown in Figure 2.

### *B. Based on OpenPose Network Structure Design*

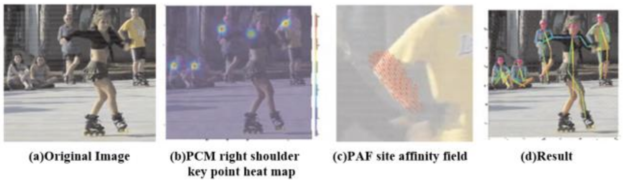
#### *1) PCM and PAF*

The crucial points of the human body can be expressed by a heat map, which is simulated using a Gaussian model, on the basis of which the values of each point represent the probability of a key point in the data of that point.

Part confidence map (PCM): This method is used to represent the Gaussian response asscociated with a pixel on the joint point. When the pixel is far away from the joint, the value of the response will increase.

Joint affinity field (part affinity fields, PAF): used to describe the spatial constraint connection between key points [14], that is, the alignment of the skeleton position and the orientation of pixels on the skeleton are crucial factors. The proximity of the predicted Part Affinity Field (PAF) to the actual PAF determines the closeness of the connection between the two nodes. Expressed by PCM there are C class vector fields, and the vector field of each limb is expressed by two feature maps, which represent the direction vectors of x and y, for a total of 19 classes, so the output of the convolutional network is 38 feature maps.

$$L^* = (L_1^*, L_2^*, ..., L_c^*), L_c^* \in R^{w\times h\times 2}, c \in \{1, ...c\} \quad (1)$$


(a)Original Image

(b)PCM right shoulder key point heat map

(c)PAF site affinity field

(d)Result

Figure 3. The process of image inspection

Basic process: As shown in Figure 3, first enter a picture (a), go through the network, get a bunch of heat maps (b) and PAF sets (c), and obtain the parsed diagram (d) after matching the dichotomous diagram.

#### *2) OpenPose network structure design*

OpenPose is a convolutional neural network-based model that undergoes enhancements for real-time, multi-person human keypoint detection within a supervised learning framework. The primary architecture of the original network is depicted in Figure 4, which is divided into two components [15]. Initially, feature extraction is accomplished through the traditional convolutional neural network VGG19 (the first 10 layers), resulting in the acquisition of feature map F. This feature map is then fed into a two-branch multi-stage network. The upper branch focuses on predicting Partial Affinity (PAF), capturing positional and directional information between key points. Simultaneously, the lower branch is dedicated to predicting a Partial Confidence Map (PCM), which characterizes the location of key points.

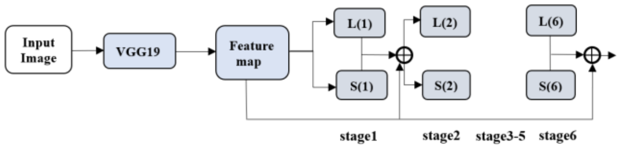


Figure 4. OpenPose's network structure diagram

The internal structure of a network for predicting partial affinity and confidence is illustrated in Figure 5. This network employs multiple stages to extract semantic information between key points.

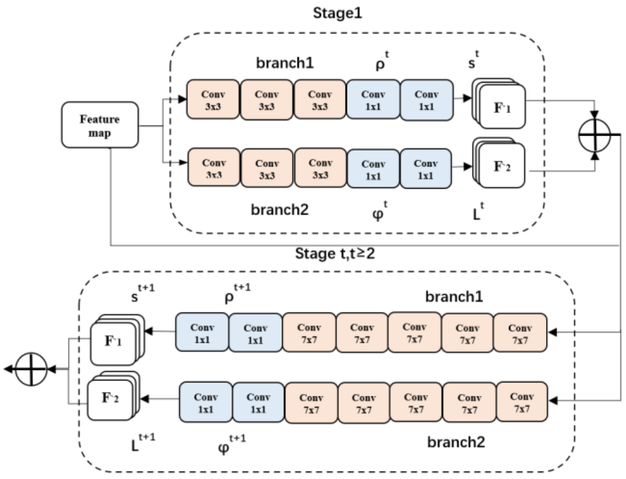


Figure 5. Diagram of the forecast internal structure

In addition to the first stage, multiple 7x7 convolution kernels are used in all other stages, which can obtain larger receptive fields [15], towards the conclusion of each stage, the forecasted values of the two subnetworks are connected with the initial characteristic curve F and used as input for the next step, In order to enhance the fusion of deep feature information without overlooking surface features, the equation description is as follows (2)(3):

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), t \geq 2 \qquad (2)$$

$$L^t = \Phi^t(F, S^{t-1}, L^{t-1}), t \geq 2 \qquad (3)$$

### C. Optimization

#### 1) Imporve feature extraction

TABLE I. LIGHTWEIGHT BACKBONE PERFORMANCE COMPARISON

| | **AP, %** |
|---|---|
| MobileNet v1<br>(cut to conv4_1) | 37.9 |
| Dilated MobileNet v1<br>(cut to conv5_5) | 42.8 |
| Dilated MobileNet v1<br>(cut to conv5_6) | 43.2 |
| Dilated MobileNet v2<br>(cut to conv6_3) | 39.6 |

Lightweight OpenPose replaces VGG as the backbone by using MobileNet_v1[16]. However, the MobileNet network structure lacks depth, as when returning to a skeletal point, attention must be given not only to the immediate vicinity but also to a broader context. This approach allows for accurate localization of the skeletal point even in the presence of occlusion, and if you simply use MobileNet, the effect is not good. Insufficient depth in the network structure hampers the attainment of a broader receptive field, consequently impacting the performance of the receptive field, so that the accuracy of bone positioning will be reduced, according to the test of convolution performance in 2D multi-person human pose estimation, as shown in Table Ⅰ, where AP represents the average accuracy and GFLOPs represent the model complexity.

Therefore, in order to improve the receptive field to get better results, you can use MobileNet (Dilated MobileNet v1) with void convolution, and the main function of hollow convolution is to expand the receptive field and obtain contextual information.

#### 2) Improve the number of branches

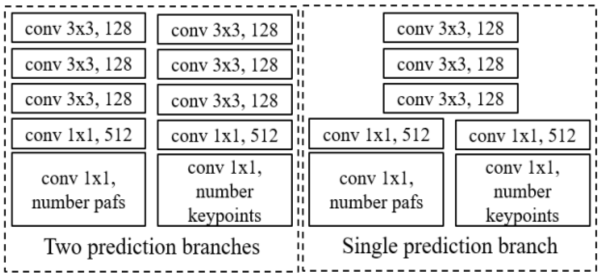


Figure 6. Merge branching diagram

The OpenPose model has two parallel branches, as shown in Figure 3.4 (left), utilized for forecasting the keypoint heatmap (keypoint map) and the keypoint affinity field (PAF). Because the two branches are the same in structure, but the number of output results is different, so lightweight OpenPose considers merging the two branches into one branch. Figure 6 (right) directly combines the original two prediction stages into one stage, and only needs to use 1*1 convolution in the last output stage to separate the two stages as output.

#### 3) Modify the convolution kernel size

Figure 4 shows the internal network structure of OpenPose, where two prediction branches are composed of multiple convolution kernels concatenated, and 7×7 convolution kernels are frequently used in the prediction network stage. Although larger convolutional kernels can obtain larger receptive fields, their computation is also large.

Therefore, on this basis, lightweight OpenPose uses 1*1, 2*3*3 size convolution cascade, opting for a small convolution kernel instead o f a larger one significantly decreases the computational workload, in order to obtain the same feeling field with 7*7 convolution kernel, add a hole convolution with an expansion parameter size of 2 to the last piece of 3*3 convolution [15][16], because the kernel of the hole convolution is not continuous, therefore, the residual connection structure is used for each piece, as shown in Figure 7.

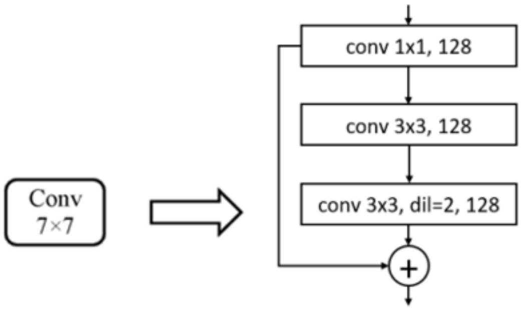


Figure 7. Improved convolution kernel structure

## IV. RESULTS AND DISCUSSION

### A. Model performance evaluation

Table Ⅱ shows the experimental results of the original OpenPose model and the improved OpenPose model under the COCO dataset, using AP value as the model evaluation index.

TABLE II. IMPROVING OPENPOSE NETWORK EVALUATION RESULTS

| Model | AP, % |
|---|---|
| OpenPose | 48.6 |
| Optimization | 86.3 |

### B. Testing on COCO Datasets

#### 1) Analysis of human bone point detection results

Since there are 3 people in the detected picture, a total of 3 people's key points are generated, and the OpenPose model has a total of 25 key points, of which 24 points are marked for Body; The improved OpenPose model in this paper uses a COCO dataset with a total of 19 points and 18 joint points, and the last point of which is labeled with the image background.

Figure 8 and Figure 9 show the confidence distribution line plot based on OpenPose model and lightweight OpenPose model to realize the key points of multi-person human posture, respectively, from which it can be seen that the distribution trend of confidence between the improved model and the original model has not changed greatly.

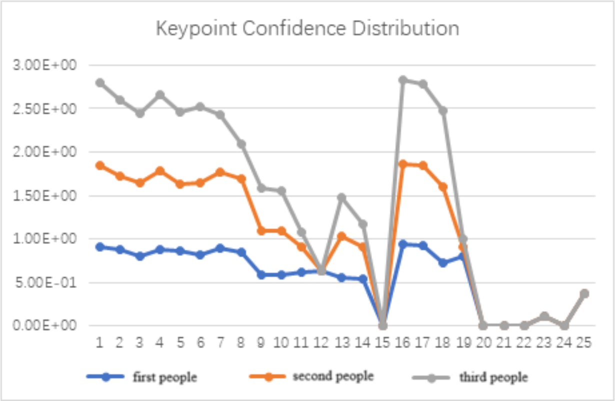


Figure 8. OpenPose model key point confidence distribution chart

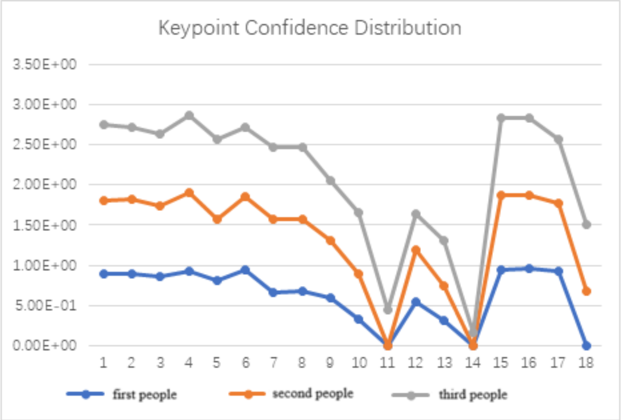


Figure 9. Lightweight OpenPose model key point confidence distribution chart

Through the confidence analysis of the key point position of the human body in Table III, the accuracy of detecting the position of the human body from OpenPose to lightweight OpenPose has no particularly large impact, and even improved, as shown in Table III.

TABLE III. COMPARISON OF AVERAGE CONFIDENCE

| Average confidence of person i | 1 | 2 | 3 |
|---|---|---|---|
| OpenPose method | 0.572 | 0.473 | 0.473 |
| Lightweight OpenPose method | 0.625 | 0.718 | 0.779 |

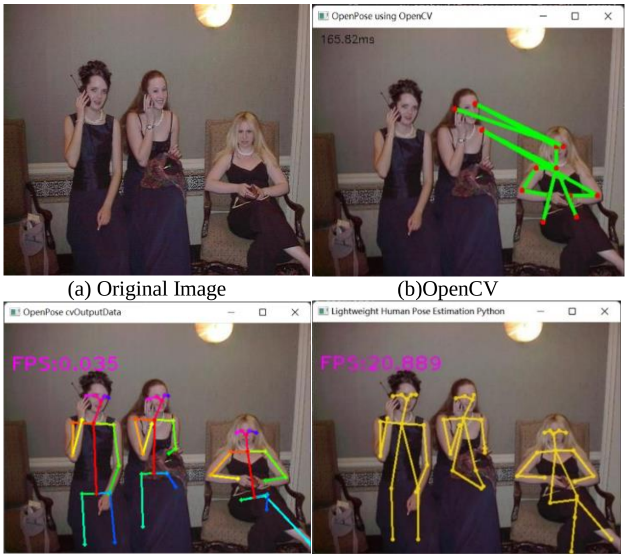
#### 2) Analysis of attitude detection rate results

The comparison of OpenPose model and improved Openpo in terms of frame number, as shown in Table Ⅳ, can assess the accelerated speed of the model proposed in this study compared to the original model, indicating the feasibility of the improved OpenPose model for estimating and recognizing human posture.

TABLE IV. COMPARISON BEFORE AND AFTER NETWORK STRUCTURE IMPROVEMENT

| Type | FPS |
|---|---|
| OpenPose | 0.035 |
| Lightweight OpenPose | 20.889 |

### 3) Multi-Person Pose Estimation Results



(a) Original Image (b)OpenCV

(c) OpenPose (d) Lightweight OpenPose

Figure 10. Multi-person pose recognition results

As shown in Figure 10, analysis of multi-person pose recognition results, (a) is the original figure, (b) is a multi-person human pose recognition image not implemented by OpenPose, compared with (c) is a multi-person human pose recognition image implemented by OpenPose, which can obviously conclude the feasibility of OpenPose to recognize multiple human postures; (d) is the improved multi-person human posture estimation, adding a human frame for each human body in the picture, (c) is compared with (d), although (c) can identify multiple human postures well, but its model is complex, and the speed is higher when implemented, and the improved model is significantly faster when testing. This implies that the improvement not only focuses on enhancing performance but also achieves a significant enhancement in the efficiency of model execution. This comprehensive optimization provides the model with stronger and more efficient support across various applications.

## V. CONCLUSIONS

Human pose estimation and recognition are widely used in modern society, for example, it is extremely convenient for people's lives, so the research on the lightweight of pose estimation has the significance of promoting the development of society and meeting the needs of people and the market. As science and technology progress, a lightweight algorithm for human pose estimation is introduced, which is more conducive to our research on pose estimation, so this paper then studies the OpenPose bottom-up convolutional neural network method proposed in 2017 and improves its model and network structure to achieve lightweight multi-person human pose estimation. According to the results realized, it is also verified that lightweight pose estimation reduces the amount of computation and reduces the complexity of the model compared with traditional OpenPose pose estimation, and is more suitable for mobile hardware devices. However, for the current mainstream lightweight human pose estimation, there are still many areas for improvement in this study:

- A splitter can be used to divide each recognized action, and each recognized action is type-output.
- When extracting features, more accurate and fast methods such as residual network ResNet18 can be used.
- You can also train your own model by collecting your own data set and apply human pose estimation to specific applications, such as: intelligent monitoring with the function of detecting falls, AI fitting to meet social fear and deep home, etc.
- Expand the study of 2D human posture to 3D research.
- Connected with software, a system can be created to detect the type of human movement.

## REFERENCES

[1] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. In IJCV, 2005.

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In CVPR, 2010.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In CVPR, 2016

[5] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In CVPR, 2016.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi person pose estimation model. In ECCV, 2016.

[7] Cao,Z,Simon,T,Wei,S,et al.Realtime multi-perpson 2d pose estimation using part affinity fields [A].// Proc of the IEEE Conference on Computer Vision and Pattern Recongnitio n [C], Honolulu, HI, USA: IEEE, 2017:1302-1310.

[8] M. Kocabas, S. Karagoz, and E. Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In ECCV, 2018.

[9] S. Hong, B. Roh, K.-H. Kim, Y. Cheon, and M. Park. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. In arXiv preprint arXiv:1611.08588, 2016.

[10] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In CVPR, 2017.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In arXiv preprint arXiv:1704.04861, 2017.

[12] B. Xiao, H. Wu, and Y. Wei. Simple Baselines for Human Pose Estimation and Tracking. In ECCV, 2018.

[13] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In CVPR, 2014.

[14] U. Iqbal and J. Gall. Multi-person pose estimation with local joint-to-person associations. In ECCV Workshops, Crowd Understanding, 2016.

[15] Li Yifan, Yuan Longjian, Wang Rui. Improved Lightweight Human Action Recognition Model Based on OpenPose % J Electronic Measurement Technology [J]. 2022, 45(01): 89-95.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.