Parallel Multi-Branch Model with Multi-Resolution Interaction for Human Pose Estimation

Lingyi Cai and Wei Liu*
School of Computer Science, Hubei University of Technology, Wuhan 430068, China
*Corresponding author: 1810300403@hbut.edu.cn

Abstract. In recent years, human pose estimation has become a research hotspot in the field of computer vision, and has extensive application value in many fields, such as activity recognition [1], action detection, self-driving, etc. This paper tries a novel multi-stage network structure to complete the task of human pose estimation, and focus on improving the loss of spatial accuracy caused by repeated downsampling and upsampling. The network structure contains multiple resolution subnets. After the original resolution is learned in the first stage, a low-resolution branch will be generated in each subsequent stage, and these branches can learn feature representations in parallel. Through continuous multi-scale and multi-resolution fusion, necessary information interaction between subnets of different resolutions can be carried out to obtain richer feature maps and make the predicted key point heat maps more accurate. This paper uses COCO keypoint detection dataset to obtain the relevant results of pose estimation, which verifies the effectiveness of this network.

Keywords: Computer Vision, Human pose estimation.

1. Introduction

The main goal of human pose estimation is to locate the pixel positions of key points (such as eyes, shoulders, elbows, etc.) in the image as accurately as possible. However, in practical applications, there are often many factors that affect the precise positioning of key points, such as occlusion, angle, lighting, complex background, various posture changes, etc. In order to solve these problems, early research used hand-made features and related models to obtain posture information. Recently, with the development of deep convolutional networks, the use of more powerful learning level representations to capture spatial context information, and reference between joint points to correct each other's errors. These effective work has made the human pose estimation breakthrough the limitations of traditional methods, so the accuracy of the forecast has also been greatly improved. 2D human pose estimation tasks are usually classified into single person pose estimation and multi-person pose estimation. The following related works can solve these two types of pose estimation tasks.

1.1. Single person pose estimation

Single person pose estimation is suitable for situations where there is only one person in the picture. Two methods are usually used: Direct regression and Heatmap based.
1.1.1. Direct regression. Direct regression method [1] uses the output feature map to directly regress the key points of the human body.

1.1.2. Detection based. The Detection based method [2-4] first generates a set of heat maps, and then predicts key points based on the heat maps. CPM [2] is a multi-stage network and Wei et al. added a convolutional network on the basis of pose machines, so it can learn image features and spatial information at the same time. Newell et al. proposed the Stacked Hourglass [3] network. It stacks multiple Hourglass modules so that each stage can use the output of the previous stage as input to obtain spatial context information, which can be used for reference between joint points and improve the prediction accuracy of joint points. It is worth noting that when the image drops from high resolution to low resolution, the spatial accuracy will be lost, and the restoration from low resolution to high resolution will make the final learned feature spatial accuracy weaker. Repeated up-sampling and down-sampling operations make it difficult to obtain accurate prediction results on tasks that are sensitive to spatial accuracy. Therefore, the design of the network structure model should not be limited to the route of recovering the high-resolution representation from the low-resolution representation generated by the convolutional neural network, but should consider establishing a new network structure for high-resolution representation learning. Based on this idea, Sun et al. proposed a network called HRNet [4], which always maintains high-resolution representations throughout the process. At the same time, the connection between the different resolution convolutions is changed from the traditional series mode to the parallel mode. This structure enables better interaction between multi-resolution representations, promotes the exchange of context and spatial information, and improves the expressive ability of high-resolution and low-resolution representations, so that you can learn enough high-resolution characterization.

1.2. Multi-person pose estimation
Multi-person pose estimation can use the network structure to predict the key points of each person when the number and position of people in the image are unknown, mutual occlusion, and crowding. It can be divided into the following two types of methods: Top-down and bottom-up.

1.2.1. Top-down. The Top-down [5-8] method first uses detectors to detect the target object, and then uses a single-person pose estimation algorithm for each bounding-box to detect key points of the human body. Chen et al. designed a two-stage network CPN [5]. The first stage GlobalNet network is a feature pyramid network based on the ResNet backbone [6], which is used to learn pyramid features to provide contextual information to locate simple key points. The second stage network RefineNet learns the characteristics of all levels and uses Online Hard Keypoints Mining to generate loss*, which effectively solves the positioning of hard keypoints (such as occlusion, complex background, etc.). However, Xiao et al. [7] believe that a simpler method can be used to obtain high-resolution images, so they introduced deconvolution to replace the structure composed of upsampling and convolution, and also achieved better results. The network architecture of MSPN [8] has two stages. The modules in each stage adopt the GlobalNet network in CPN, and use coarse-to-fine supervision to obtain a large amount of contextual spatial information to improve positioning accuracy. At the same time, the features of the same layer of different stages are merged to better solve the problem of information loss caused by repeated upsampling and downsampling.

1.2.2. Bottom-up. Bottom-up [9, 11] first predicts all the key points from the input image, and then associates the key points to the corresponding person. The Deepecut [9] method first use Fast R-CNN [10] to obtain an integer linear program to perform NMS on a group of body-part candidates generated using CNN variants, then group and assemble them, and finally combine with the fully connected model to complete the task of human pose estimation. Openpose [11] is a multi-stage network architecture, each of which follows the structure of CPM [2]. It uses two branches to process feature maps. The first branch is used to generate a discrete set of body candidate parts, the second branch
encodes the position and direction information of the area supported by the limbs. The prediction of each branch is combined with the original image features as the next stage. Finally, it use greedy inference to generate Bipartite Matching to match the candidate body parts to the corresponding person, and output the key points of everyone in the image.

1.3. The proposal of this paper
Some of above methods to convert the input image from high resolution to low resolution, and then restore from low resolution to high resolution. This kind of network structure may lose part of the spatial position information, resulting in weaker accuracy of the finally learned feature space. However, it seems difficult for this kind of network structure to repeatedly stack in series to obtain a large performance improvement. This article believes that a relatively different multi-stage network structure, namely Parallel multi-branch model with multi-resolution interaction for human pose estimation, can be used to effectively alleviate the above-mentioned problems and further improve the accuracy of prediction.

The Bottom-up method uses a direct regression method and is simple and effective, but the prediction effect for dense crowds is not ideal. The Top-down method does not match the joint points incorrectly, and its detection accuracy is high, but it takes a long time. This article uses the Top-down method to estimate the pose of a single human body using the model proposed in this article for each object in the generated bounding box. The specific method is to design each different resolution representation as a separate branch, and each branches are parallel to the other. At the end of each stage of the network, the new branch and the high-resolution branch will be residually connected, so that the feature information between different resolutions is fully integrated, which can effectively alleviate the problem of spatial information loss caused by frequent down sampling and upsampling in some of the above methods, and further accurate estimate spatially heatmap to improve the accuracy of key point detection and the efficiency of related parameters.

2. Method
The network structure is divided into 4 stages. The first stage uses 4 residual modules to process the original input image. Each subsequent stage will generate a new branch with a resolution reduced by half, and these branches will perform deep residual learning in this stage. Then, the new branch and the old branch are respectively exchanged for information at the end of this stage, and the final feature representation obtained will be used as the input of the next stage. In order to not only extract high-level features, but also effectively retain the original level of information, a residual module is used in the entire network and stacked appropriately to better obtain the cross-level information interaction.

2.1. Residual module in this paper
Generally speaking, as the number of network layers increases, problems such as gradient disappearance (explosion) and degradation will be caused, and some spatial information will be lost during convolution and other processes. The proposal of residual module [6] can effectively improve these problems. Residual module consists of residual representations and identity shortcut connections, as shown in Figure 1. Its output can be expressed by the formula: \( y = S(\mathcal{F}, \{W_i\}) + W_{c} \mathcal{F}. \) \( S(\mathcal{F}, \{W_i\}) \) represents the process of learning residual mapping. The shortcut connection part (i.e., \( \mathcal{F} \), see Fig 1.) skips two or more layers and represents the process of performing identity mapping. \( W_{c} \) is used to match the shapes of \( \mathcal{F} \) and \( y \) when necessary.
2.2. Multi-stage network structure with parallel multi-branches

2.2.1. Multi-stage network structure. The network structure of this article is divided into 4 stages. Each branch maintains the same resolution with deep residual learning. Even if the network is deep, it is not easy to lose some important spatial context information, so it can extract feature maps more effectively with the stacked residual modules.

Let $F_b^t$ be the feature map in the $t$th stage and $b$ be the $b_{th}$ branch where $t, s \in \{1, 2, 3, 4\}$. The initial input image is denoted as $F_1^0$, and a set of feature maps $F_b^1$ is generated after the convolutional network of the branch where the first stage is located. In the following three stages, each stage will generate a new branch with smaller resolution feature maps. Without considering the branch information interaction, the input of each stage is the output of the previous stage, so the feature map $F_b^t$ of each stage can be expressed as: $F_b^t = \mathbb{S}_b(F_b^{t-1})$, $b \leq t$ where $\mathbb{S}_b$ represents the deep residual learning of $b_{th}$ branch in stage $t$.

2.2.2. Multi-branch parallel structure. It is mentioned in section 3.2.1 that when $b>1$, a branch with a smaller resolution will be generated at each stage, so the strided $3 \times 3$ convolution with the stride 2 was adopted, which is denoted as $\mathbb{R}_b$ in this paper. Specifically, the initial feature map of the branch $b$ generated in the stage $t$ is $F_b^t = \mathbb{R}_b(F_b^{t-1})$, $b > t$. Therefore, each time a new branch is generated, its resolution is reduced to half of the original and its width (The number of channels) is increased to the double.

2.3. Information exchange between branches of different resolutions

In the parallel process of multi-resolution branches, this article directly uses downsample to obtain new branches, which may also cause the problem of information loss. In order to allow lower resolution branches to learn more comprehensive spatial context information, they need the exchange of information among multiple branches, while also reducing the difficulty of training. Therefore, a multi-resolution information interaction module is set in the network structure, which can be regarded as an extended residual design. When a new branch is generated at a certain stage, each branch will first perform deep residual learning, and then merge the new branch with a smaller resolution with the upper branch for cross-branch feature fusion. At the same time, in order to improve network efficiency and reduce network load, branches with higher resolution will not have too much interaction in this module. In this paper, this method is recorded as the function $\mathbb{M}_b$, which means that effectively merges the feature representation of branch c and branch b. Due to the frequent and variability of information interaction, the residual module in this module is diverse. The method used in this article is residual connection to preserve the spatial context information as best as possible. It also can effectively correct the key points of the predicted error. For low-resolution feature representation, this article uses
appropriate deconvolution to restore it to the corresponding higher resolution, which can learn more rich feature sets that are a robust image representation for both the analysis and synthesis of images.

For example, in the third-stage information interaction module in the network, the down-sampling function $M_3$ uses two consecutive $3 \times 3$ convolutions with the stride 2 for 4X downsampling and one strided $6 \times 6$ convolution with the stride 4 for shortcut connection. The recovery high resolution (upsampling) function $M_3$ uses two consecutive $4 \times 4$ deconvolutions with the stride=2 for 4X upsampling and one $4 \times 4$ deconvolution with the stride 4 for shortcut connection.

Therefore, only considering the feature fusion of each branch at each stage, the resulting feature representation $F_{b}^{(3)}$ is:

\[
F_{b}^{(3)} = \sum_{i=1}^{t-1} M_{b}^{i}(S_{b}^{i}(F_{b}^{i})), \quad b=t
\]

\[
F_{b}^{(3)} = M_{b}^{b+1}(S_{b}^{(t-1)}), \quad b<t
\]

2.4. Network instantiation

Combining the design of the above three parts, this article uses the following formula to describe the characteristics of each branch in each stage:

\[
F_{b}^{(t)} = \begin{cases} 
S_{b}^{t}(F_{b}^{t-1}) + M_{b}^{b+1}(S_{b}^{t}(F_{b}^{t-1})) & b < t \\
S_{b}^{t}(F_{b}^{t-1}) + \sum_{i=1}^{t-1} M_{b}^{i}(S_{b}^{i}(F_{b}^{i})) & b = t \\
N_{b}^{t}(F_{b}^{t-1}) & b > t 
\end{cases}
\]

Each branch in each stage of this network structure uses 4 residual modules and corresponding shortcut connections, and cross-branch spatial feature information exchanges were performed 2, 4, and 6 times in the second, third, and fourth stages. On the basis of the network structure of this article, you can also try to add more stages and generate more branches to explore the accuracy of the final prediction.

![Figure 2. Illustrating the structure of the proposed network in this paper. It consists of 4 high-to-low resolution branches in parallel. The newly generated branch will interact with the high-resolution branches (feature fusion).](image)

3. Experiments

The COCO dataset contains images of 91 objects types and 250,000 person instances labeled with 17 keypoints. The network model proposed in this article was trained on COCO train2017 and evaluated
on val2017 set and test-dev2017 set. The evaluation indicator used is object keypoint similarity (OKS) which is defined as:

\[
\text{OKS} = \frac{\sum_{i} \exp \left( -\frac{d_i^2}{2s_i^2}\right) \delta(\|v\| > 0)}{\sum_{i} \delta(\|v\| > 0)}
\]

During training, the widths of the 4 branches of the network structure used in the experiment are 32, 64, 128, and 256 respectively. The size of the input image is reset to 256×192, the data is appropriately expanded, and the basic learning rate is set to 1e-3. In the end, a total of 200 epochs were trained. In the test, a two-stage process similar to CPN is used: first, the human detector is used to detect human instances, and then the location of each key point is predicted. Table 1 shows the results obtained on val2017 set. Compared with the CPN network, the AP of this article has a gain of 7.2. Compared with the SimpleBaseline network, the results obtained in this article have improved in all indicators. Table 2 shows the attitude estimation performance of the method in this paper and the previous excellent technologies on the test-dev set. The AP of the method proposed in this paper reaches 74.9, and all indicators are better than CPN and SimpleBaseline network. The results in Table 1 and Table 2 fully demonstrate the effectiveness of the network structure proposed in this paper. In the experiment, the ablation study of the information interaction between different branches in the network was carried out. (a) means that the first branch and the fourth branch are interacted during the whole process; (b) means that the second branch and the fourth branch are interacted during the whole process; (c) means that during the whole process Interact the third branch with the fourth branch. This paper adopts the structure (3), and the results are shown in Table 3, which proves the effectiveness of the different resolution information interaction part of the network proposed in this paper.

### Table 1. Comparisons on the COCO validation set.

| Method             | Backbone | Input size | AP  | AP^{50} | AP^{75} | AP^{M} | AP^{L} | AR  |
|--------------------|----------|------------|-----|---------|---------|--------|--------|-----|
| CPN [5]            | ResNet   | 256×192    | 68.6| -       | -       | -      | -      | -   |
| Simple Baselines [7] | ResNet   | 384×288    | 72  | 89.3    | 79.8    | 68.7   | 78.9   | 77.8|
| This paper         | This paper | 384×288    | 75.5| 90.3    | 82.5    | 71.7   | 71.8   | 80.8|

### Table 2. Comparisons on the COCO test set.

| Method             | Backbone | Input size | AP  | AP^{50} | AP^{75} | AP^{M} | AP^{L} | AR  |
|--------------------|----------|------------|-----|---------|---------|--------|--------|-----|
| CPN [5]            | ResNet   | 384×288    | 72.1| 91.4    | 80.0    | 68.7   | 77.2   | 78.5|
| Simple Baselines [7] | ResNet   | 384×288    | 73.7| 91.9    | 81.1    | 70.3   | 80.3   | 79  |
| This paper         | This paper | 384×288    | 74.8| 92.3    | 82.7    | 71.2   | 80.5   | 80.1|

### Table 3. Ablation study.

| Method | (a) | (b) | (c) | AP  |
|--------|-----|-----|-----|-----|
| (1)    | ✓   |     |     | 70.3|
| (2)    | ✓   | ✓   |     | 71.6|
| (3)    | ✓   | ✓   | ✓   | 73.3|

### 4. Conclusion

The network structure tried in this article contains multiple branches. Each branch subnet represents a different resolution. These subnets will exchange information at each stage, and effectively solve the spatio-temporal information missing problem caused by repeated downsampling and upsampling. Finally, a set of accurate key point heat maps can be generated for human pose estimation, but the disadvantage of this network is that it takes a long time to predict the key points. Future work can improve the network in this paper and apply it to other research fields such as face detection and semantic segmentation.
References

[1] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multitask deep learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[2] Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016.

[3] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European conference on computer vision. Springer, Cham, 2016.

[4] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[5] Chen, Yilun, et al. "Cascaded pyramid network for multi-person pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Xiao, Bin, Haiping Wu, and Yichen Wei. "Simple baselines for human pose estimation and tracking." Proceedings of the European conference on computer vision (ECCV). 2018.

[8] Li, Wenbo, et al. "Rethinking on multi-stage networks for human pose estimation." arXiv preprint arXiv: 1901.00148 (2019).

[9] Pishchulin, Leonid, et al. "Deepcut: Joint subset partition and labeling for multi person pose estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.