High-resolution Human Pose Estimation Based on Location Awareness

WU Ning[a], Hui GAO*[b], Peng WANG[c], LI Xiaoyan[b], LV Zhigang[b]
[a]School of Ordnance Science and Technology, Xi'an Technological University, Xi'an , China 710021; [b]School of Electronics and Information Engineering, Xi'an Technological University, Xi'an, China 710021; [c]Development Planning Office, Xi'an Technological University, Xi'an, China 710021
* Corresponding author: 976834875@qq.com

ABSTRACT

In crowded and complex scenes, it is easy to cause problems such as poor human pose estimation and low key point positioning accuracy. In this paper, a high-resolution human pose estimation algorithm based on position awareness was proposed. The algorithm introduced the coordination attention (CA) in the feature extraction module, which realized the accurate acquisition of the spatial position information of key points, finally improved the human pose. Estimate the detection accuracy of the algorithm. The AP value of the improved algorithm was 76.5%, which was 2.1% higher than the original algorithm, the AP^{50} was increased by about 3.1%, and the AP^{75} was increased by about 2.8%. The experimental results showed that the proposed algorithm could effectively improved the detection performance in crowded and complex backgrounds, and had higher detection accuracy.

Keywords: Human Pose Estimation, High Resolution Network, Coordination Attention

1. INTRODUCTION

With the continuous development of computer vision field, human pose estimation as its branch had also been widely developed and applied. The field of human pose estimation had been widely used in various fields of human social life. With the continuous development of virtual reality, action recognition, medical rehabilitation and other fields, higher requirements were also placed on the accuracy of human pose estimation, and the development was in the ascendant [1].

In recent years, with the introduction of deep convolutional neural networks, the field of human pose estimation had developed rapidly. Human pose estimation based on deep learning could be divided into two types [2] , one was the bottom-up framework, which directly locates the key points of the human skeleton, and then matches the key points to the corresponding people through clustering and other methods [3]. Another was the top-down framework, which first detected the human body, and then located and recognized key points through regression methods or heatmap methods [4-6]. The literature [7] proposed the Cascaded Pyramid Network CPN (Cascaded Pyramid Network), which effectively improved the detection accuracy of occlusion key points by using online hard example mining. The literature [8] proposed the High-resolution network (HRNet), high-resolution representation effectively improved the accuracy of human pose estimation network. The literature [9] proposed a scale-aware high-resolution network (Higher HRNet), It added multi-scale fusion to the high-resolution network, which improved the challenge of scale transformation. The literature [10] proposed a lightweight high-resolution network (Lite-HRNet), in the high-resolution network introduced a conditional channel weighting unit to replace the expensive point wise convolution, which effectively reduced the complexity of the network.

Although the high-resolution network HRNet had better detection accuracy, the detection effect of human pose estimation in crowded and complex scenes was not ideal. Therefore, this paper started the research with the high-resolution network as the backbone network, where coordinate attention CA was introduced to enrich the keypoints location information and enhance the spatial information acquisition capability of the network, so as to improve the detection accuracy of the human pose estimation algorithm in complex scenes.

Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022), edited by Xianye Ben, Proc. of SPIE Vol. 12462, 124620I · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2660942

2. HIGH RESOLUTION NETWORK

The high-resolution network HRNet always maintains a higher resolution in the entire network structure, and connects sub-networks from high resolution to low resolution in parallel. HRNet had four stages, with the high-resolution subnetwork as the first stage, and gradually added subnetworks from higher to lower resolutions to form a new stage, followed by parallel subnetworks contains the resolution subnetworks from the previous stage, and newly generated lower resolution subnetworks. The high-resolution network used parallel multi-resolution sub-networks to extract semantic information, while serial multi-resolution sub-networks fuse features, and continuous multi-scale features to fuse. HRNet contains four parallel subnetworks, and a schematic diagram of the high-resolution network is shown in Figure 1.

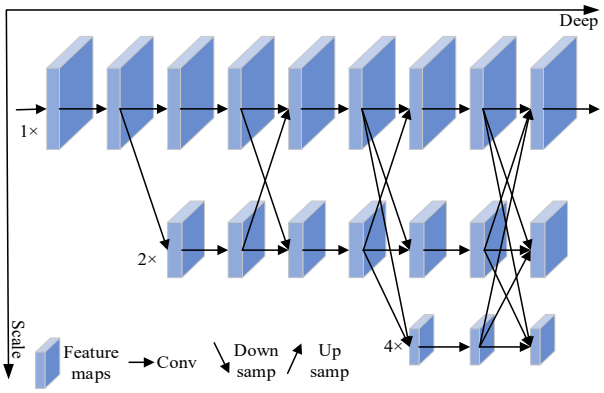


Figure 1. High-resolution network architecture

3. IMPROVED HIGH RESOLUTION NETWORK

3.1 CA Module

The CA module of the coordinate attention mechanism was shown in Figure 2. When any intermediate feature tensor was input, C was the number of channels, H was the height of the input feature, and W was the width of the input feature. The CA module divides the input into a one-dimensional feature encoding process in two directions, one of which captures long-range dependencies in spatial directions , the other captures precise location information in spatial directions. The resulting feature maps were encoded into a pair of orientation-aware and position-sensitive feature maps. Respectively, so that the input feature map information could be supplemented to enhance the representation of keypoint locations. The final output transform tensor with precise position information had the same size as the input tensor X.

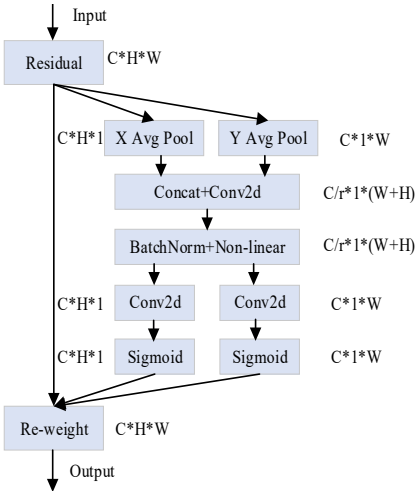


Figure 2. Coordinate attention

3.2 C-Basicblock Module

Aiming at the problem that key point detection in complex background was easily affected by the surrounding environment, in order to enhance the extraction of key points location information in complex environment by HRNet feature extraction network. This paper introduced the location attention mechanism CA (Cooraditation Attention) module. The information was embedded into the channel attention, which enhanced the feature extraction network ability to obtain spatial coordinate information in complex backgrounds, which not only enables the network to obtain rich high-dimensional feature information but also obtains accurate location information. Thereby, improving the subsequent keypoints positioning and prediction effect , to enhance the robustness of the network. In the feature extraction part of the HRNet network, multiple coordinate attention CA modules were integrated to construct a C-Basicblock residual module for location feature extraction. The overall structure of the reconstructed feature extraction module C-Basicblock was shown in Figure 3.

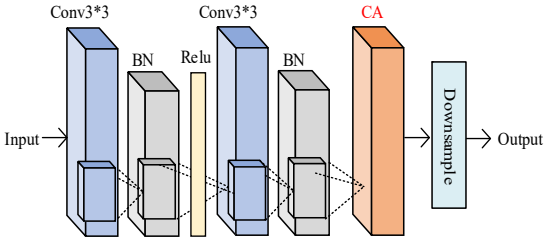


Figure 3. Residual block C-Basicblock

3.3 CA-HRNet

In order to improve the human pose estimation effect of the network in complex scenes, under the framework of deep learning, this paper used the high-resolution network HRNet as the basic network, and used the constructed C-Basicblock module to reconstruct HRNet, and finally obtained a high-resolution location-aware network. The high-resolution human pose estimation network CA-HRNet, and its overall network results were shown in Figure 4.

Given an input data, the CA-HRNet network first preprocessed it, and then used the Bottleneck module, C-Basicblock module, Up Sample and Down sample operations of the high-resolution network to ensure sufficient extraction of features at different resolutions.And fully obtained the location information of key points, and finally performed feature fusion on the extracted relevant information to achieve accurate regression positioning of key points.

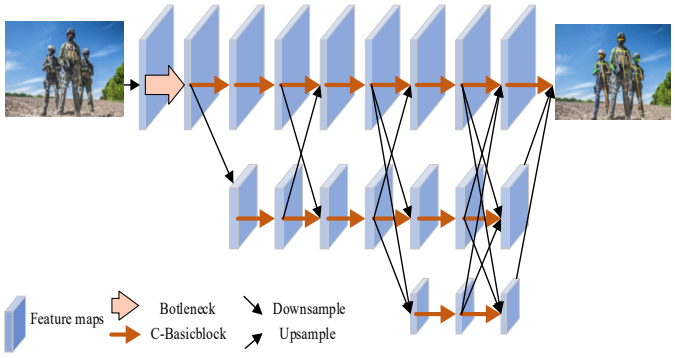


Figure 4. Human pose estimation network CA-HRNet

4. EXPERIMENTAL RESULTS AND ANALYSIS

The algorithm in this paper was based on the Pytorch deep learning framework, the experimental operating system was Ubuntu 18.04, and the server hardware configuration was Nvidia GeForce RTX 2060Ti graphics card. In this experiment, the Adam optimizer was used to optimize the model. All experiments in this paper were carried out in the above environment.

4.1 Image Preprocessing

In the model training process, the image was first preprocessed. The main preprocessing operations include cropping, flipping, and scale transformation. The cropping aspect ratio of the image was 256:196. The random flip strategy was used to randomly rotate the image (- 45°~ + 45°) and change the image scale, the scale change was 0.7:1:1.35 three different scales. In the model training process, the Adam optimization algorithm was used to update the network weights iteratively. After every 3.6 million iterations, the learning rate was reduced by two times. The initial learning rate was 5e-4, the batch size was set to 16, and the training period was set to 210 batches.

4.2 Dataset

The dataset used for model training in this paper was the MS COCO2017 dataset, which included a total of 57K images and 150K human instances, included 64115 training samples and 5000 validation samples, which were divided into training sets, validation sets, and test set as needed.Section headings.

4.3 Evaluation Indicators

In the MS COCO dataset, OKS (Object Keypoint Similarity) was used to represent the target keypoint similarity, whose value was distributed between (0,1), and the better prediction result when the value of OKS was close to 1. The formula for OKS was shown in Equation (1).

$$OKS = \frac{\sum_i \frac{-d_i^2}{e^{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i [\delta(v_i > 0)]} \tag{1}$$

Where d_i represents the euclidean distance between each predicted key point and the real key point, which v_i means that the key point was visible, s represents the object scale, and k_i represents the control attenuation constant.

In order to evaluate the performance of the human pose estimation model, this paper adopted the main evaluation indicators of MS COCO data set AP, AP^{50}, AP^{75}, AP^{M}, AP^{L}, AR, AR^{50}, AR^{75}, AR^{M}, AR^{L}. Where AP value indicates the average precision at 10 OKS thresholds (OKS=0.50, 0.55, 0.60, ..., 0.95), AP^{50} indicates the detection precision at OKS=0.50, AP^{75} indicates the detection precision at OKS=0.75, AP^{M} indicates medium The detection accuracy of objects, AP^{L} represents the detection accuracy of large objects. AR represents the average recall rate at 10 OKS thresholds, and each value of AR had the same meaning as AP.The predicted key point position and the marked key point position were calculated by formula (1) to calculate the OKS value, and then the corresponding AP and AR values could be calculated to obtain the final detection precision and recall rate.

4.4 Experimental Results and Analysis

4.4.1 Model Training

Figure 5 showed the accuracy change curve of the improved model training in this paper. It could be seen from the figure that when the model started training to about 165 batches, the model adaptively adjusted the learning rate, so that the accuracy increased and the change gradually became stable.

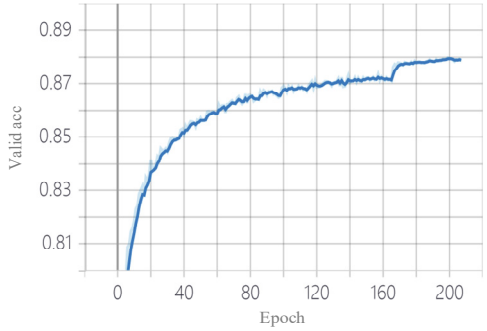


Figure 5. Model training accuracy

4.4.2 Quantitative Analysis

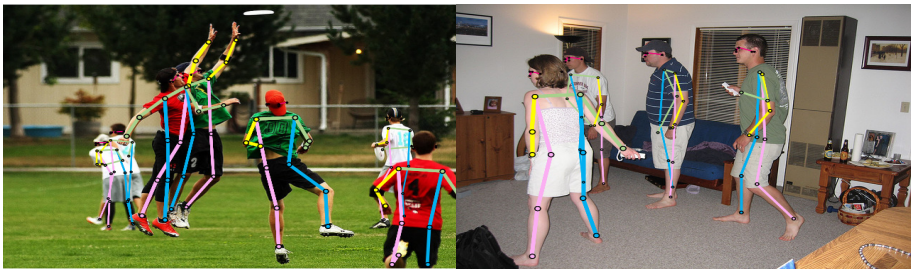
In order to verify the effectiveness of the algorithm in this paper, experiments were carried out on the COCO2017 data set. The experimental results were shown in Table 1. As can be obtained from the table1, compared with the benchmark model HRNet, the AP value of the algorithm CA-HRNet in this paper had increased by about 2.1%, and the AP^{50} had increased by about 3.1%, and AP^{75} had increased by about 2.8%. Therefore, it could be obtained that the CA-HRNet model effectively improved the network's ability to acquire key point location information, enhanced the localization effect of human key points in complex scenes, and thus effectively improved the detection accuracy of the human pose estimation algorithm.

Table.1 Performance comparison of different models

Model	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
HRNet	74.4	90.5	81.9	70.8	81.0	79.8	94.2	86.5	75.7	85.8
CA-HRNet	76.5	93.6	84.7	73.8	82.3	80.1	94.4	86.1	76.2	86.9

4.4.3 Visual Analysis

The verification results of the algorithm in this paper on the COCO data set was shown in Figure 6. In Figures (a) and (b), when occlusion occurred in a complex environment, the model could ensure the detection of key points occluding the human body, and fully extracted the information of occlusion key points, so as to realized the positioning of key points of the human body. In Figure (c), when there were branches in the background that were similar in shape to the limbs, the model in this paper could to use the contextual semantic information and the model to infer the location of the occlusion key points according to the logical positions of the skeleton joints, which effectively avoids the influence of the branches. In Figures (d), (e), and (f), the crowded and complex scenes made it difficult to locate key points, but the model in this paper could still fully extract the spatial location information, so as to achieve accurate positioning in crowded and complex scenes. This verified the effectiveness of the algorithm in this paper.



(a) (b)

(c) (d)

(e) (f)

Figure 6. Validation results on COCO dataset

5. CONCLUSION

In order to improve the detection effect of the model in crowded and complex scenes, this paper proposed a high-resolution human pose estimation algorithm based on position awareness. The coordinate attention mechanism was introduced to improve the high-resolution network, and multiple CA modules were used to enhance the location information acquisition of the model in complex scenes, enriched the contextual semantic information, and improved the detection accuracy of human key points. Experimental results showed that the algorithm in this paper could fully obtain keypoints location information, enhance the effect of key point localisation in crowded and complex scenes, and improve the accuracy of the human pose estimation algorithm.

ACKNOWLEDGMENT

This paper is one of the stage achievements of the projects supported by the National Natural Science Foundation of China (62171360), the Scientific Research Program Funded by Shaanxi Science and Technology Department (2022GY-110), and the Xi'an Technological University President's Fund Surface Cultivation Project (XGPY200217).

REFERENCES

[1] Chen W, Yu C, Tu C, et al. A survey on hand pose estimation with wearable sensors and computer-vision-based methods[J]. Sensors, 2020, 20(4): 1074.

[2] Munea T L, Jembre Y Z, Weldegebriel H T, et al. The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation[J]. IEEE Access, 2020, 8: 133330-133348.

[3] Kreiss S, Bertoni L, Alahi A. Pifpaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 11977-11986.

[4] Angelini F, Fu Z , Long Y, et al. 2D Pose-Based Real-Time Human Action Recognition With Occlusion-Handling[J]. IEEE Transactions on Multimedia, 2020, 22(6): 1433-1446.

[5] Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation[C]//Proceedings of the 2016 European Conference on Computer Vision. Springer, Cham, 2016: 483-499.

[6] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[7] Chen Y L, Wang C H, Peng Y X, et al. Cascaded Pyramid Network for Multi-Person Pose Estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, UT, USA: IEEE, 2018: 7103-7102.

[8] K Sun, Xiao B, Liu D , et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5693-5703.

[9] Cheng B W, Xiao B, Wang J D, et al. Higher HRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation[C]//Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, WA, USA: IEEE, 2020: 5385-5394.

[10] Yu C Q, Xiao B, Gao C X, et al. Lite-HRNet: A Lightweight High-Resolution Network[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 10435-10445.