Signal, Image and Video Processing (2022) 16:551–558
https://doi.org/10.1007/s11760-021-01999-y

ORIGINAL PAPER

TSNet : Tree structure network for human pose estimation

TianJun Wan[1,2] · YanMin Luo[1,2] · Zhiqian Zhang[1,2] · Zhilong Ou[1,2]

Received: 23 February 2021 / Revised: 27 July 2021 / Accepted: 29 July 2021 / Published online: 11 August 2021

Abstract
Multi-person pose estimation in natural scenes has been a hot topic in the recent years. The prediction speed of the top-down methods is affected by the number of people in the scene, so the bottom-up methods has an advantage in natural scenes. However, the study found that the accuracy of human margin joints (the joints farther from the center of the human, such as wrist and ankle) is always lower than that of the joints that are closer to the center of the human (such as shoulder and hip), and the accuracy gap between joints categories is large. Inspiring from the structural characteristics of human body, this paper proposes a tree structure network (TSNet) for human pose estimation, which divides the joints of the human into several levels according to the characteristics of human body structure, and stepwise predicts the joints from human center to human margin. Combining with the global features, the joint features of the next layer are predicted by extracting the correlation between the joint features of the current layer and the joint features of the previous layer. Therefore, each human joint contains not only the joint information of the current layer and the joint information of the previous layer, but also the background information. The experiment results show that this method can effectively alleviate the uneven precision of joints, and the TSNet can effectively improve the accuracy of lower body joints by setting different activation values for different joints. Extensive experiments on MPII datasets demonstrate the effectiveness of our proposed model and method.

Keywords Tree structure network · Different activation values · Hierarchy of joint division · Multi-person pose estimation

1 Introduction

Human pose estimation is a fundamental task of computer vision to the study of human behavior. It allows the computer to detect the position of human joints from a single RGB image. Due to the requirements of actual scenes, multi-person pose estimation has been more popular than single-person pose estimation in the recent years.

Currently, human pose estimation methods can be categorized into top-down methods and bottom-up methods. The top-down methods [1–4] depend on a person detectors to detect each person instance with a bounding box. Then perform the single person pose estimation [6–9] for each person instance. Generally, the top-down methods can achieve high detection accuracy. Most of state-of-the-art methods on multi-person human pose estimation benchmarks are based on top-down methods. However, those methods has high computational complexity and lacks the correlation between various joints of the body and global context information of other people. Because of the lack of end-to-end design, the prediction results of the joints in the second stage will be affected by the results of person detectors in the previous stage. Moreover, the top-down methods are easily affected by the number of people in the image, the detection speed of those methods will decreases with the increase in the number of people. By contrast, the bottom-up methods [11–13,15] directly detects the candidate joints of all the people in the image, than categorize the candidate joints into the corresponding human instance according to artificial defined clues between the joints. This strategy eliminates the limitation on the number of people in the image. It can run with high efficiency and is more capable of achieving real-time performance. Although the bottom-up methods has made sig-

✉ YanMin Luo
lym@hqu.edu.cn

TianJun Wan
wantj@stu.hqu.edu.cn

Zhiqian Zhang
zqzhang@stu.hqu.edu.cn

Zhilong Ou
19013083010@stu.hqu.edu.cn

1 College of Computer Science and Technology, Huaqiao University, Xiamen 361021, PR China

2 Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, PR China

nificant progress on common datasets, it is still challenged by complex postures and real-world scenarios.

Extensive research [5,10,14,16,17] has shown that the accuracy of the human margin joints is always much lower than that of other joints. We notice that the accuracy always decreases gradually from the shoulder joint, the elbow joint to the wrist joint. The accuracy gap between different joints is large. The same result goes from hip to ankle. This phenomenon also occurs in the top-down methods, but the accuracy gap of which is smaller. There are three main reasons for the smaller accuracy gap: (1) Comparing with other joints, human margin joints are more flexible and similar. (2) When it comes to the occluded joint points, it is very difficult to detect them. (3) During training, the occluded human marginal joints were not well trained. Therefore, it is significant to improve the accuracy of human margin joints.

In this paper, we propose a new tree structure network to solve the above problem of multi-person pose estimation. We assume that the feature information between adjacent joint points is more important for joint point prediction than the joint itself and some background features, especially when the joint points are covered. In addition, in predicting phase, the margin joints with higher degree of freedom, a more average activation value is needed.

In summary, our contributions are as follows:

1. We tried to solve the problem of low precision of human edge joints, which has rarely been studied before in bottom-up multi-person pose estimation.
2. We propose a TSNet for multi-person attitude estimation based on the bottom-up process, in which joints are categorized into different levels. During the training, the joint features of the current layer were combined with the global background features to predict the next layer of joint one by one. It can make full use of the feature extraction ability of neural network for each subtask, also increase the amount of the context information between joints and promote the information flow between the joints of the same layer. Consequently, the prediction accuracy of edge joints is improved.
3. We set different activation values for different joints so that the joints with a higher degree of freedom can learn to have a data which is distributed more averagely. To verify the validity of the network on the same dataset, we designed experiments with TSNet of two different structures and replaced the network of other methods with TSNet of our best by using the same processing operations and matching rules.

The remainder of this paper is organized as follows. Section 2 introduces some related work of human pose estimation. Section 3 discusses the proposed method. The two different network structures are introduced in Section 3.1 and Section 3.2, respectively. Section 4 analyzes the experimental results. Section 5 summarizes the effectiveness and limitations of the proposed method and looks forward to the future work.

2 Related work

Top-down methods. In the top-down methods [2,27,34], the bounding box of each person is first detected in the image, and then a single-person pose estimation is performed for each bounding box. The person bounding boxes are usually generated by an object detector [19–22]. Most previous methods used well-trained and state-of-the-art (SOTA) human detectors, such as Faster R-CNN [21] and SSD [23]. Mask R-CNN [24] directly adds a keypoint detection branch on Faster R-CNN. The faster and more accurate YOLO [25] is widely used at present. There are two options for single-person pose estimation. One is represented by the DeepPose [30]. It is different from the traditional method [26–28] of matching handcrafted skeletal features. DeepPose adopts a cascade method, uses a deep network model to transform human pose estimation into the keypoint regression process. The other [18] option is heatmap, which can directly detect human joints. Numerous networks are designed for feature extraction based on the above two methods [29], following that the accuracy is greatly improved.

Bottom-up methods. Conversely, in the bottom-up methods [32,33], all joint candidates are first detected by applying a joint detector globally, and then the joints are categorized into multiple human instances through the clues of artificially defined joints. In terms of joint detector design, to accurately detect the high flexibility and small scale of the marginal joints in the human body, Stacked Hourglass Network [31], the HigherHRNet [36], and the method [28,35] similar to Stacked Hourglass Network. These methods can extract and integrate high-resolution appearance features and low-resolution semantic features, obtaining more accurate detection results. For the matching phase, the offset between human joints is usually used to build the model. This method is represented by OpenPose [14], which defines the affinity of joints through the geometric offset of human torsos. PPN [17] works by mapping candidate joints to different human embedding centers, which can capture more contexts to alleviate the problem of the joints' being obscured. On the contrary, Chen et al. [38] proposes a method of predicting limb heatmaps as cues between the joints to enhance the characteristic information between the joints and improve the accuracy of the joints. To eliminate the matching phase and improve the prediction speed of human posture estimation, SMPM [37] directly detects the center of each human and gets other joint positions through joint coordinate regression.

3 Our method

In this section, we will discuss about tree structure network. Figures 1 and 2 illustrate two kinds of designed network architecture. To begin with, the first method will be briefly described. Then, according to the result analysis of the first method, the other method and its components will be introduced in detail.

3.1 Initial TSNet

Compared to the previous work, we did not predict all the joint heatmaps and all the offset heatmaps in the corresponding branches. Instead, we consider each joint type as a branch of the network. The network framework is shown in Fig.1a. There are 16 branches, namely 16 joint types, in the network. Each branch contains a feature module responsible for predicting a joint heatmap and an offset heatmap of a joint. The thorax joint was chosen to be the root node of the human, where the output feature map of each feature module is combined with the global background feature map to predict the feature of the next adjacent joint. The information transfer process between branches of the network is shown in Fig.1c. All the joint heatmaps and offset heatmaps predicted by the branches are connected in the end, and the final output of the two parts is obtained by making slight adjustments through a convolution block, respectively. This method is out of the consideration of the structural characteristics of human body, so that more contextual information can be captured for the edge joint with a high degree of freedom.

Keypoint heatmaps. Both methods use the same generation approach for joint heatmaps that adopt the form of Gaussian response heatmap used in most algorithms. The calculation process of the keypoint heatmap is shown in Equation (1). Take the joint p_j as an example, $C_{j,p}$ denotes the generated confidence score of the joint in the heatmap position p. σ_j denotes an empirical constant set to control the variance of the Gaussian distribution, usually, this value is an empirical fixed value, and if the distance value between p and p_j is bigger than the set threshold d, then the confidence score for this position is set to 0, otherwise, equation 1. Eventually, a heatmap will produce a collection of the same type of joints. We choose the maximum confidence to be the ground truth for the joint, which is the position of a joint.

$$C_{j,p} = exp(\frac{-\|p_j - p\|^2}{\sigma_j^2}) \qquad \|p_j - p\|^2 \leq d \tag{1}$$

Different from the previous methods, the proposed algorithm sets different values σ_j for different types of joints to control the activation values generated in the joint heatmap. Table 1 shows the results are affected by different values.

Offset heatmaps. Both methods use the same generation approach for the clue between the joints. We use the form of offset embedding similar to PPN. Specifically, first, we define a joint as the root node of the human and choose the thorax joint. Then, we offset each connected joint from the position of the back joint to that of the front one. The offset value is also the position corresponding to the maximum confidence in the joint heatmap. When the distance value between two kinds of joints is smaller than the set threshold, the two joints are adjacent joints. $F^i_{jx\tau}$ and $F^i_{jy\tau}$ denote the response value of the position of the j^{th} joint of the i^{th} people on the $2j^{th}$ and $(2j+1)^{th}$ offset heatmaps, respectively. The response values of offset heatmap are calculated according to Equations (2) and (3), where w and h are, respectively, the width and the height of the input image in the training stage. If the distance value between p^i_{jx} or p^i_{jy} and τ is bigger than the set threshold d, then the confidence score for this position is set to 0, otherwise, Equation 2 and equation 3.

$$F^i_{jy\tau} = \frac{p^i_{jy} - \tau_y}{h} \qquad \|p^i_{jy} - \tau_y\|^2 \leq d \tag{2}$$

$$F^i_{jx\tau} = \frac{p^i_{jx} - \tau_x}{w} \qquad \|p^i_{jx} - \tau_x\|^2 \leq d \tag{3}$$

During the experiment, a jump connection of joint features was added to enhance the context information between the joints, as shown by the red arrow in Fig.1c. By adding the jump connection can improve the accuracy of human edge joints. But it is easy for a redundant structure to transmit the prediction error of the front joint to the back joint, the error is magnified at the end joint. And we found something else. As shown in Table 1, by setting different activation values for different joints, the accuracy of leg joints can be improved effectively.

3.2 Tree structure network

TSNet improves the network structure on the basis of Initial TSNet, the result of TSNet is better. As shown in Fig. 2 (b), the global features were extracted through an hourglass module. The thorax joint is considered as the root node of the human. The network will learn joint heatmaps and offset heatmaps separately. Specifically, we divide the human joints into four layers according to their distance from the center of the human. The first layer contains the upper neck, chest and pelvis; the second layer contains the top of the head, shoulders and hips; the third layer contains the elbows and knees; and the fourth layer contains the wrists and ankles. Each layer is a branch of the network, and the joint characteristics of each layer are predicted separately. The joint features of the next layer are predicted by combining the fea-

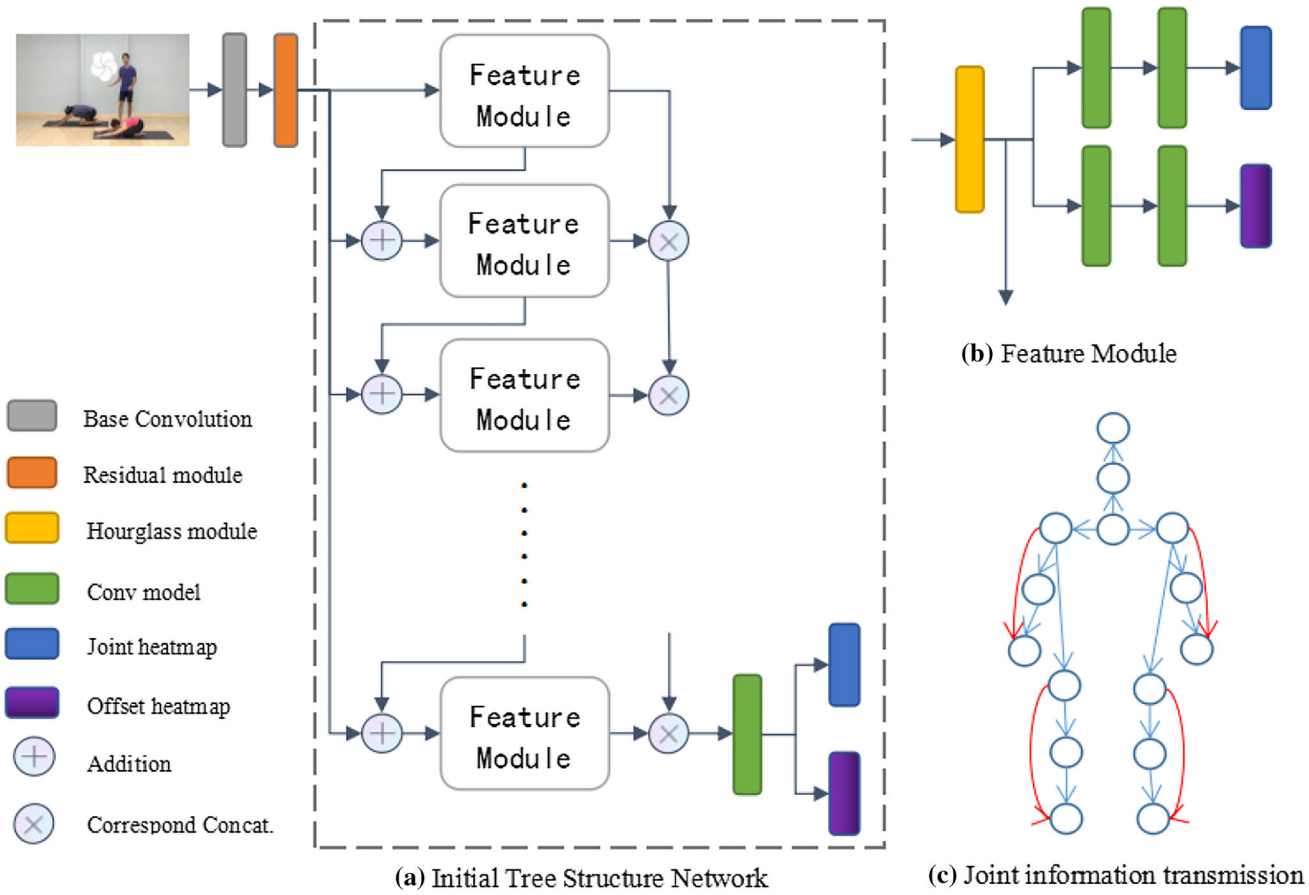

Fig. 1 a The overall structure of initial TSNet has 16 branches, each of which contains a feature module, that is, there are 16 feature modules in total. b The structure of the feature module contains three outputs, namely joint heatmap, offset heatmap and joint feature map. c The process of joint information transmission in the network. The joint features are successively transferred from the human chest joint to the next joint. The red line represents the added transmission in the experiment

Table 1 Ablation study of Initial TSNet by setting the Gaussian variance for different joints on MPII dataset

σ	Head	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Total
fixed value	91.4	87.6	75.4	64.2	65.9	58.2	50.1	70.4
gradually dec.	90.4	86.6	74.6	62.2	66.6	59.1	47.6	69.6
gradually inc.	91.2	86.7	75.5	63.9	68.5	62.3	52.2	71.5

tures of the previous layer and the background features. The offset heatmaps are obtained in the same way.

The network framework is shown in Fig.2a. The design of the network imitates [31] in the way of network stacking. First, the image is scaled to a basic scale according to what size of a person is given in the dataset. The image size is (W,H,C), where W is width, H is height, and C is the number of channels. A basic feature graph F (has 256 channels) is obtained by a basic convolution block and a residual block. Firstly, the feature graph F is input into an hourglass module to obtain the background feature F*. The output results of each branch in each TSNet are polymerized with the background feature F*. Then, the polymerized results are put into the next TSNet to predict again. In the experiment, we stacked eight.

3.3 Loss function

Through the modeling method of the above task, the output results of each layer branch in each TSN are integrated to obtain two parts of output: joint heatmap and migration heatmap. There are N numbers of outputs in the network. The total loss includes the loss of each TSNet output. During training, The respective loss of joint heatmap and offset heatmap of each TSNet can be expressed by Equations (4) and (5). Here T is the number of layers that we divide the

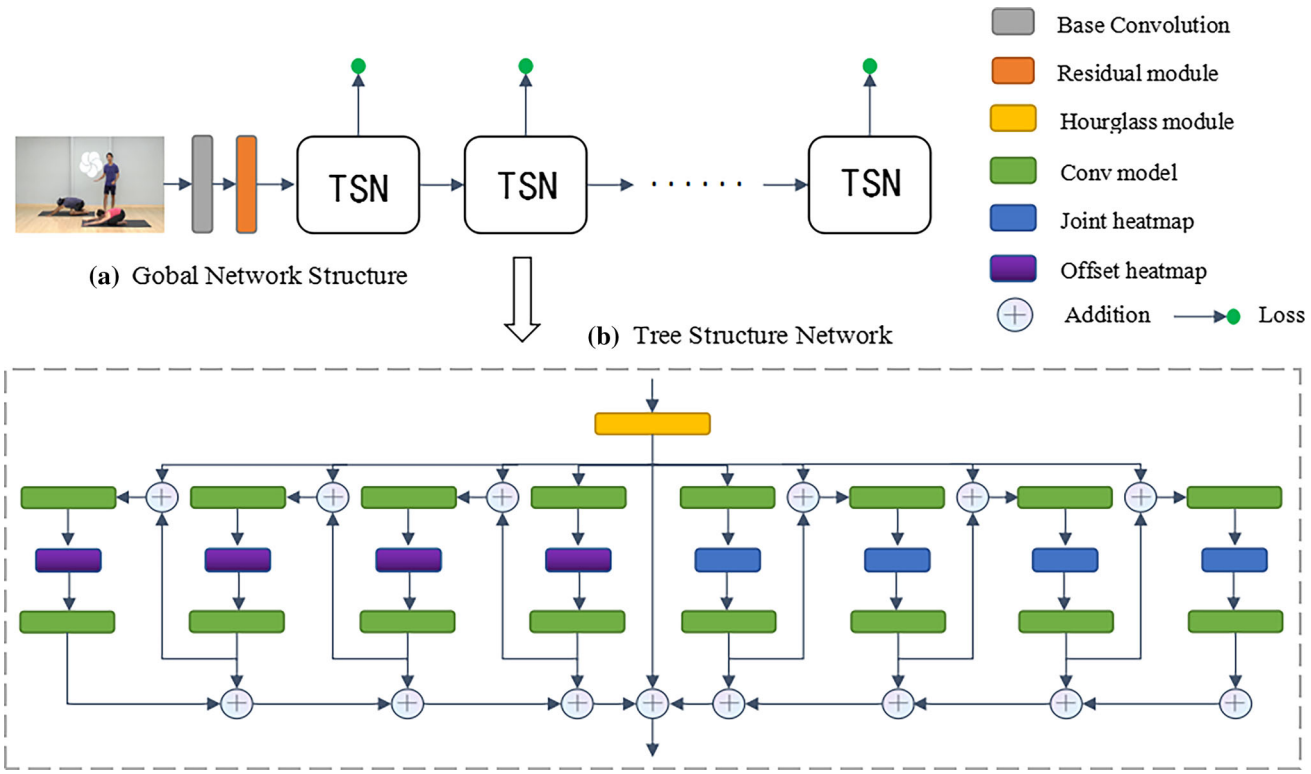

Fig. 2 a The overall structure of the network that composed of multiple TSNets cascaded. b The tree structure network adopts the method of multilayer prediction step by step

human joints. In the t layer, $S_t(P)$ and $S_t^*(P)$ denote the value of the position P in the joint heatmap and the ground truth. $L_t(P)$ and $L_t^*(P)$ are the values of the position P in the offset heatmap and the ground truth. The total loss can be calculated by formula (6), where λ_1 and λ_2 represent the weight of each part of the loss function, respectively. (λ_1=0.5,λ_2=0.5)

$$Loss_S = \sum_{t=1}^{T} \sum_{P} \|S_t(P) - S_t^*(P)\|_2^2 \tag{4}$$

$$Loss_L = \sum_{t=1}^{T} \sum_{P} \|L_t(P) - L_t^*(P)\|_2^2 \tag{5}$$

$$Loss_{total} = \sum_{stage=1}^{N} (\lambda_1 Loss_S + \lambda_2 Loss L) \tag{6}$$

4 Experiments

4.1 Datasets and evaluation metrics

MPII Dataset. The MPII Human Pose dataset [42] consists of images taken from a wide range of real-world activities with full-body pose annotations. The MPII dataset is still a very challenging public dataset for the tasks of human pose estimation. There are around 25K images with 40K subjects with annotated body joints, which include 12K images for the training.

Training. In order to get better robustness, we use scaling and rotation methods are used to adapt to people of different scales in the image. Like many of the previous methods, we use data augmentation with random rotation ([30°, 30°]), random scale ([0.8, 1.2]) and random translation ([-40, +40]) to crop an input image patch with the size of 256×256 as well as randomly flip for comparing with other methods.

Testing. We selected 350 images from the validation set for verification. The model is trained for a total of 100 epochs in the rest of the training image samples. The initial learning rate is set to 3e-3. The model is implemented with PyTorch and the RMSProp is adopted for optimization. In the test, we crop image by using the given position and average person scale of test images, and the cropped samples are resized and padded to 384×384 as input to TSNet. Some visualization results of the images are shown in Fig.3.

4.2 Ablation experiments

From the results of the Initial TSNet, it was observed that if the joint was covered, the position of the posterior joint might not be accurately predicted. To avoid this problem, we added a jump connection of joint features. As shown in Table

Table 2 Comparison with or without jump connections, which are used by con. means

Net	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Init_TSNet	89.4	86.4	74.9	62.7	63.9	58.0	49.4	69.2
Init_TSNet+con.	91.4	87.6	75.4	64.2	65.9	58.2	50.1	70.4

Table 3 Comparison results on the MPII validation set

Net	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Iqbal et al.	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
Insafutdinov et al.	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Levinkov et al.	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Xiao Chen et al.	90.9	86.2	71.5	58.3	70.0	63.3	55.3	70.8
Insafutdinov et al.	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao et al.	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Ours(Initial_TSNet)	91.2	86.7	75.5	63.9	68.5	62.3	52.2	71.5
Ours(TSNet)	92.0	88.7	77.2	65.5	73.1	67.0	56.9	74.3

Fig. 3 Some visualization results of our approach on the benchmark of the MPII dataset

2, there was a significant improvement in all the accuracy of joints. We find that different types of joints should have different characteristics. We set different Gaussian variances for different joints so that the edge joints with a high degree of freedom can learn different activation values. As shown in Table 1, there was nearly a 3% improvement in the accuracy of the lower body joints.

Under the same algorithm, the accuracy of TSNet increases by 3% compared with that of Initial TSNet as shown in Table 3. Compared with other methods, our method can reach a relatively close result and has better results in some joints, which proves the effectiveness of the network model and method.

In order to further prove the effectiveness of the network structure, eliminate the influence of the algorithm itself on the results. We fully adopted the data enhancement, matching algorithm and training parameters used by the PPN [17] model to compare the TSNet model with the PPN model. PPN uses a much larger scale and rotation range for data enhancement than we do. In the joint matching algorithm, all the joints of each person are gathered together for cluster analysis, and in the process of prediction, multi-scale method was used to verify the results. We train and test on the same dataset. We trained with one NVIDIA GeForce GTX-2080Ti GPU and one CPU Intel I9-9900K 3.6GHz. As shown in

Table 4 Comparison of results of different networks under the same algorithm

Net	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
PPN*	92.2	89.7	82.1	74.4	78.6	76.4	69.3	80.4
Our(TSNet)	93.0	90.1	80.3	75.1	79.5	73.5	69.9	80.2

*Data published in the original paper

Table 4, under the same dataset and algorithm, the accuracy of the head, shoulder, and hip was improved. It is worth mentioning that the accuracy of human edge joints, such as wrist and ankle, was improved by 0.7%, which fully demonstrated the effectiveness of the method.

5 Conclusion

In this work, we focus on improving the accuracy of the human edge joints and propose a TSNet network to simulate the structure of human body. We observed that TSNet can set activation values of different sizes according to different joints, which can effectively improve the precision of edge joints. We modified the network structure and divided the joints into different levels, which could increase the information flow between the layers. Experimental results show that the method is effective. Under the same algorithm, the accuracy of the human edge joints can be effectively improved. Because we focus on solving the accuracy of the human edge joints and do not pay much attention to the running speed of the network, it is relatively slow in terms of speed, which is also a common problem in human pose estimation tasks. In the future, we will do some work on network quantization and network pruning, so as to reduce the number of model parameters and accelerate the prediction speed while ensuring the accuracy.

Acknowledgements This work was supported by Natural Science Foundation of Fujian Province, China under grant 2020J01082, and in part by the Science and Technology Bureau of Quanzhou under Grant 2018C113R, and in part by the National Natural Science Foundation of China under Grant 61901183

Declarations

Funding The Natural Science Foundation of Fujian Province, China under grant 2020J01082, and in part by The Science and Technology Bureau of Quanzhou under Grant 2018C113R, and in part by the National Natural Science Foundation of China under Grant 61901183.

Conflicts of interest There are no conflicts of interest.

Availability of data and material The data comes from the common dataset

Code availability Custom code

References

1. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. Mur phy, K.: Towards Accurate Multi-person Pose Estimation in the Wild, in: Proceedings of the CVPR, (2017), pp. 3711-3719
2. Fang, H., Xie, S., Tai, Y., Lu, C.: RMPE: Regional Multi-person Pose Estimation, in:Proceedings of the ICCV, (2017), pp. 2353-2362
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded Pyramid Network for Multi-person Pose Estimation, in: Proceedings of the CVPR, (2018), pp. 7103-7112
4. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J.: R ethinking on Multi-Stage Networks for Human Pose Estimation, CoRR abs/1901.0 0148 (2019)
5. Su, K., Yu, D., Xu, Z., Geng, X., Wang, C.: Multi-Person Pose Estimation With Enha nced Channel-Wise and Spatial Information, in: Proceedings of the CVPR, (2019), pp. 5667-5675
6. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional Pose Machines, in: Proceedings of the CVPR, 2016, pp. 4724-4732
7. Liang, S., Sun, X., Wei, Y.: Compositional Human Pose Regression, in: Proceedings of the ICCV, (2017), pp. 2621-2630
8. Liu, W., Chen, J., Li, C., Qian, C., Chu, X., Hu, X.: A Cascaded Inception of Inception Network With Attention Modulated Feature Fusion for Human Pose Estimation, in: Proceedings of the AAAI, (2018), pp. 7170-7177
9. Tang, W., Yu, P., Wu, Y.: Deeply Learned Compositional Models for Human Pose Estimation, in: Proceedings of the ECCV, (2018), pp. 197-214
10. Ke, L., Chang, M.-C., Qi, H., Lyu, S.: Multi-Scale Structure-Aware Network for Hu man Pose Estimation, in: Proceedings of the ECCV, (2018), pp. 731-746
11. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, in: Proceedings of the CVPR, (2016), pp. 4929-4937
12. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model, in: Proceedings of the ECCV, (2016), pp. 34-50
13. Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B.: ArtTrack: Articulated Multi-Person Tracking in the Wild, in: Proceedings of the CVPR, (2017), pp. 1293-1301
14. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields, in: Proceedings of the CVPR, (2017), pp. 1302-1310
15. Newell, A., Huang, Z., Deng, J.: Associative Embedding: End-to-End Learning for Joint Detection and Grouping, in: Proceedings of the NIPS, (2017), pp. 2274–2284
16. Kreiss, S., Bertoni, L., Alahi, A.: PifPaf: Composite Fields for Human Pose Estimation, in: Proceedings of the CVPR, (2019), pp. 11969-11978
17. Nie, X., Feng, J., Xing, J., Yan, S.: Pose Partition Networks for Multi-person Pose Estimation, in: Proceedings of the ECCV, (2018), pp. 705-720
18. XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking, in: Proceedings of the European conference on computer vision (ECCV). (2018): 466–481
19. Cheng, Bowen., Wei, Yunchao., Shi, Honghui., Feris, Rogerio., Xiong, Jinjun., Huang, Thomas.: Decoupled classification20refinement: Hard false positive suppression for object detection. arXiv preprint arXiv:1810.04002, (2018). 2

20. Cheng, Bowen., Wei, Yunchao., Shi, Honghui., Feris, Rogerio., Xiong, Jinjun., Huang, Thomas.: Revisiting rcnn: On awakening the classification power of faster rcnn. In ECCV, (2018).2
21. Ren, Shaoqing., He, Kaiming., Girshick, Ross., Sun, Jian.: Faster r-cnn: Towards real-time object detection with region proposal networks. In NeurIPS, (2015). 2
22. Lin, Tsung-Yi., Doll'ar, Piotr, Girshick, Ross, He, Kaiming, Hariharan, Bharath, Belongie, Serge: Feature pyramid networks for object detection. CVPR **2**(3), 5 (2017)
23. Liu, Wei., Anguelov, Dragomir., Erhan, Dumitru., Szegedy, Christian., Reed, Scott., Fu, Cheng-Yang., CBerg, Alexander.: Ssd: Single shot multibox detector. In ECCV, (2016). 3
24. He, Kaiming., Gkioxari, Georgia., Doll'ar, Piotr., Girshick, Ross.: Mask r-cnn. In ICCV, (2017)
25. Redmon J., Divvala, S., Girshick, R., et al.: You Only Look Once: Unified, Real-Time Object Detection[C]// Computer Vision & Pattern Recognition. IEEE, (2016)
26. Andriluka, M.., Roth, S.., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation, in: Proceedings of the CVPR, (2009), pp. 1014-1021
27. Sun, M., Kohli, P., Shotton, J.: Conditional regression forests for human pose estimation, in: Proceedings of the CVPR, (2012), pp. 3394-3401
28. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet Conditioned Pictorial Structures, in: Proceedings of the CVPR, (2013), pp. 588-595
29. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation, in: Proceedings of the CVPR, (2019), pp. 5686-5696
30. oshev, A. T., Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Net works, in: Proceedings of the CVPR, (2014), pp. 1653-1660
31. Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation, in: Proceedings of the ECCV, (2016), pp. 483-499
32. Papandreou, George., Zhu, Tyler., Chen, Liang chieh., Gidaris, Spyros., Tompson, Jonathan., Murphy, Kevin.: Personlab: Person pose estimation and instance segmentation with a part-based geometric embedding model. In ECCV, (2018).1, 2, 5, 6
33. ZHU, X., JIANG, Y., LUO, Z.: Multi-person pose estimation for posetrack with enhanced part affinity fields[C]//ICCV PoseTrack Workshop. (2017), 7
34. ZHANG, H., OUYANG, H., LIU, S.: ff. Human pose estimation with spatial contextual information[J]. arXiv preprint arXiv:1901.01760, (2019)
35. Luo, Y., Xu, Z., Liu, P., Du, Y., Guo, J.: Multi-Person Pose Estimation via Multi-Layer Fractal Network and Joints Kinship Pattern. TIP **28**, 142–155 (2019)
36. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T., Zhang, L.: HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation, in: Proceedings of the CVPR, (2020), pp. 5386-5395
37. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-Stage Multi-Person Pose Machines, in: Proceedings of the ICCV, (2019), pp. 6950-6959
38. Chen, X., Yang, G.: Multi-Person Pose Estimation with LIMB Detection Heatmaps[C]// 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, (2018)
39. Zhang, F., Zhu, X., Dai, H., et al.: Distribution-aware coordinate representation for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020): 7093-7102
40. Zhang, Zhiqian, Luo, Yanmin, Gou, Jin: Double anchor embedding for accurate multi-person 2D pose estimation[J]. Image and Vision Computing **111**(1), 104198 (2021)
41. Ou, Zhilong., Luo, YanMin., Chen, Jin., Chen, Geng.: SRFNet: selective receptive field network for human pose estimation.J Supercomputing (2021). https://doi.org/10.1007/s11227-021-03889-z
42. BULAT, A., TZIMIROPOULOS, G.: Human pose estimation via convolutional part heatmap regression[C]//European Conference on Computer Vision. Springer, (2016): 717–732