Multi-Residual Module Stacked Hourglass Networks for Human Pose Estimation

Wenxia Bao✉, Yaping Yang, Dong Liang and Ming Zhu

(Key Laboratory Intelligent Computing and Signal Processing of the Ministry of Education, Anhui University, Hefei 230039, China)

Abstract: A multi-residual module stacked hourglass network (MRSH) was proposed to improve the accuracy and robustness of human body pose estimation. The network uses multiple hourglass sub-networks and three new residual modules. In the hourglass sub-network, the large receptive field residual module (LRFRM) and the multi-scale residual module (MSRM) are first used to learn the spatial relationship between features and body parts at various scales. Only the improved residual module (IRM) is used when the resolution is minimized. The final network uses four stacked hourglass sub-networks, with intermediate supervision at the end of each hourglass, repeating high-low (from high resolution to low resolution) and low-high (from low resolution to high resolution) learning. The network was tested on the public datasets of Leeds sports poses (LSP) and MPII human pose. The experimental results show that the proposed network has better performance in human pose estimation.

Key words: human pose estimation; residual learning; image feature; hourglass network

CLC number: TP 391 **Document code**: A **Article ID**: 1004-0579(2020)01-0110-10

Human pose estimation aims to determine the precise pixel position of human body joints from an input image. It is a fundamental yet challenging task in computer vision and an important basis for high-level vision tasks such as behavior recognition, human-computer interaction, clothing parsing, re-identification and motion capture. An effective human pose estimation must be able to deal with large pose changes, changes in clothing and lighting conditions, severe body deformation and body occlusion problems. One of the key issues in solving these problems is how to extract powerful low-level and mid-level appearance features to capture relevant contextual information on each scale, and how to establish complex part relationships to achieve effective pose inference.

Received 2018-11-16

Supported by the National Natural Science Foundation of China (61401001; 61501003; 61672032)

✉ Author for correspondence, associate professor, Ph. D.

E-mail: bwxia@ahu.edu.cn

DOI: 10.15918/j.jbit1004-0579.18151

Traditional methods for human pose estimation are based on pictorial structure models, which use tree structures to model the spatial relationship of rigid body parts, usually using a predefined human body model assuming that each body part is independent of all other body parts except those connected to it. These models can describe pose accurately, but their main disadvantage is that they need to hand-design the structure of the model to capture specific dependencies between different output variables while allowing for easy inference.

Methods based on a pictorial structure model are limited by feature expression ability. In practical applications, the flexible movement of the human body and the constant changes in vision and clothing lead to a large change in the visual information of body parts, thus bringing great challenges to human pose estimation. Human pose estimation has recently showed significant progress with the widespread use of convolutional

neural networks (ConvNets) and many auxiliary methods, such as resnet, inception design and batch normalization.

Human pose estimation based on ConvNets can be classified as detection-based or regression-based. Detection-based methods use ConvNets as part detectors, then combine the graphical model to estimate pose, which requires the manual design of the energy function or the heuristic spatial probability prior initialization to remove outliers on the regressive score map. Some of these methods also use private networks for precise correction. For example, Tompson et al. [1] used multiple branches of ConvNets to fuse features of image pyramids and used Markov random fields (MRF) for post processing. This architecture can improve the performance of pose estimation using the geometric relationship between the part positions. Chu et al. [2] proposed a geometric transformation kernel for the convolutional layer of ConvNets, which can effectively characterize the dependence between keypoints, and a bidirectional tree structure model, which can make each keypoint receive information from all relevant joints and optimize their features. Bulat et al. [3] designed a detection-regression ConvNet cascade network structure, which first uses the detection sub-network to output the score maps of each part and then uses the regression sub-network for joint regression. The regression-based method mainly uses ConvNets to realize the nonlinear mapping from the image to human pose. Pfister et al. [4] proposed a spatial fusion layer and combined the optical flow information from multiple frames in the video to capture the dependencies between joints, so as to estimate the human pose in the video. Carreira et al. [5] use an iterative error feedback network that introduces top-down feedback on the standard model and iteratively feeds the learned pose information back to the first layer of the network to gradually correct the initial prediction.

Since human pose estimation based on ConvNets have proven to have significant advantages, how to design a dedicated network structure with higher accuracy and robustness for human pose estimation has become a new research direction. Therefore, based on the regression-based approach, a human pose estimation method based on multi-residual module stacked hourglass networks (MRSH) is proposed in this paper. Three different residual modules are used in the network as the building blocks for an hourglass sub-network. When the resolution is high, the large receptive field residual module (LRFRM) and the multi-scale residual module (MSRM) are used to capture information in a larger range and on each scale. When the resolution is reduced to the lowest level, only the improved residual module (IRM) is used, and intermediate supervision is performed at the end of each hourglass sub-network, repeating the high-low, low-high learning. The final network in this paper uses four multi-residual module hourglass sub-networks and verifies network performance on two standard public datasets.

1 Hourglass Sub-Network

1.1 Common hourglass sub-network

Fig. 1 is a common hourglass sub-network consisting of pool, upsample and residual modules. The hourglass network is designed to capture information at every scale in a feed-forward fashion. Local features are essential for learning smaller parts such as arms and legs, and the final pose estimation requires a coherent understanding of the whole body. For example, the orientation of the person, the position of the limbs, and the relationship between adjacent joints are some of the most recognizable cues at different scales in the image. The hourglass network captures these features better and combines them to ultimately output pixel predictions.

During the learning process of the feature map, the size of the feature map of the input hourglass sub-network is 64×64. First, the fea-

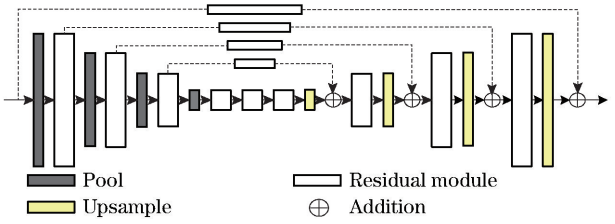

Fig. 1 Schematic diagram of the common hourglass sub-networks

ture map is downsampled to perform high-low processing. The network achieves the lowest resolution at 4 pixels and then performs low-high processing by upsampling the features from the bottom layer combined with the higher resolution features.

1.2 Residual module

The building block used by the common hourglass sub-network is the residual module shown in Fig. 2. The module consists of two branches: the first branch is the identity map branch and the second branch is composed of three convolution layers (the size of the convolution kernel is 1×1, 3×3, 1×1), batch normalization, and relu. The use of this residual module not only improves the performance of the ConvNets, but also reduces the training burden on the network.

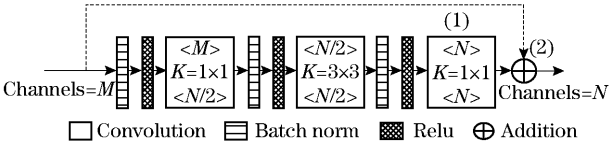

Fig. 2 Schematic diagram of residual module

2 Multi-Residual Module Stacked Hourglass Network

2.1 Improved residual module

Residual learning[6-7] uses identity mapping to enable us to train very deep neural networks, but it is also the root cause of residual learning: identity mapping increases the response change when the network is deeper, thereby increasing optimization difficulty.

Therefore, we use batch normalization, rrelu and 1×1 convolution to replace the identity mapping in the previous method (the branch (1) in Fig. 3). The advantage of this substitution is that

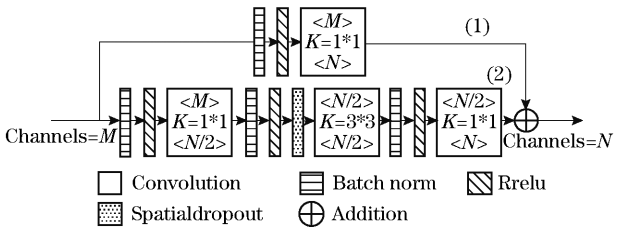

Fig. 3 Schematic diagram of IRM

when the resolution of the feature map is reduced, or the dimension of the feature channel is increased, the identity mapping is replaced by the convolution layer, and the network can reset the response variance to a smaller value, thereby effectively avoiding larger variance response.

In addition, we introduce a spatial dropout layer[8] and use rrelu activation to the IRM before the convolution in the middle of the branch (2) of the original residual module. The role of the dropout layer is to improve the generalization ability of the model by preventing the activation from becoming strong and preventing over-fitting. Due to the full-convolution network we use, there is a strong spatial correlation between the joints of human body on the training image, and the activation of the feature map is also has strong correlation. As a result, the ordinary dropout layer cannot regularize its output; otherwise it will lead to a significant decline in learning rate. In this case, the spatial dropout proposed in Ref. [8] can help the network learn the correlation between adjacent pixels on the feature map.

2.2 Large receptive field residual module

We focus on human pose estimation, in which larger receptive fields[9-10] have been shown to be important for locating local body parts, and the context area of neurons is its corresponding receptive field. Although the IRM proposed has a certain improvement on the performance of the model, the receptive field of this module is small. In order to better learn the correlation between the joints of the human body, a large receptive field residual module is designed on the basis of the IRM.

Fig. 4 shows a schematic diagram of the LR-

FRM, which is mainly composed of convolution, batch normalization, activation layer, pool, upsample, and spatial dropout layers. This module adds a branch to the IRM, that is, the branch(3) in Fig. 4, which includes two 3×3 convolution layers, each convolution layer has a batch normalization and an activation layer before it, and the spatial dropout layer is added before the second 3×3 convolution layer. The addition of this convolution branch can effectively expand the receptive field of the output layer.

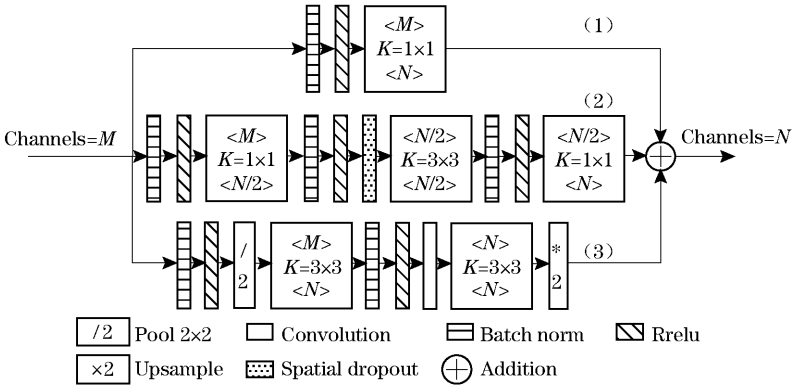

Fig. 4 Schematic diagram of LRFRM

Deep residual networks have made great breakthroughs in image classification tasks by introducing stacked residual modules, which can be represented as

$$p_{i+1} = h(p_i) + F(p_i, W_i^F) \quad (1)$$

where p_i and p_{i+1} are the input and output of the i-th residual module, F is the stacked convolution, batch normalization and rrelu nonlinearity, and $h(p_i) = p_i$ is the identity mapping.

The expression of LRFRM designed in this paper is

$$p_{i+1} = h(p_i) + F(p_i, W_i^F) + Q(p_i, W_i^Q) \quad (2)$$

Each LRFRM consists of three branches. Branch(1) has only one 1×1 convolutional layer, and the receptive field of the output layer is 1×1. Branch(2) is $F(p_i, W_i^F)$, a group of convolutions (conv1 $\times 1$ + conv3 $\times 3$ + conv1 $\times 1$), so the receptive field of the output layer is 3×3. Branch(3) is $Q(p_i, W_i^Q)$, with structure pool2 $\times 2$ + conv3 $\times 3$ + conv3 $\times 3$ + upsample2 $\times 2$. Due to the pool, the resolution for convolution in this branch is half of the other two branches, and the receptive field of the output layer is 10×10. These three branches have different receptive fields, so this module increases the effective receptive field through branch(3), while using branches(1) and (2) to retain the high-resolution information of the original residual module, and finally fusing into the next module.

2.3 Multi-scale residual module

Based on the LRFRM, a MSRM is proposed, which consists of convolutional, batch normalization, activation layer, pool, upsample and spatialdropout layers. As shown in Fig. 5, the module adds branch(4) to the LRFRM, including a 1×1 convolution and two 3×3 convolutions, in which the first two convolution layers have a downsampling, and the last 3×3 convolution layer is preceded by a spatial dropout layer. In front of each convolution layer there is a batch normalization and an activation layer. The representation of this module is

$$p_{i+1} = h(p_i) + F(p_i, W_i^F) + Q(p_i, W_i^Q) + G(p_i, W_i^G) \quad (3)$$

The MSRM consists of four branches. The branches(1)(2)(3) are consistent with the LRFRM, and branch(4) is the new design in this paper. Its structure is pool2 $\times 2$ + conv3 $\times 3$ + pool2 $\times 2$ + conv3 $\times 3$ + upsample2 $\times 2$ + upsample2 $\times 2$. Branch (4) is downsampled twice, so the feature map is reduced to one-fourth of the input feature map and then merged with the other three bran-

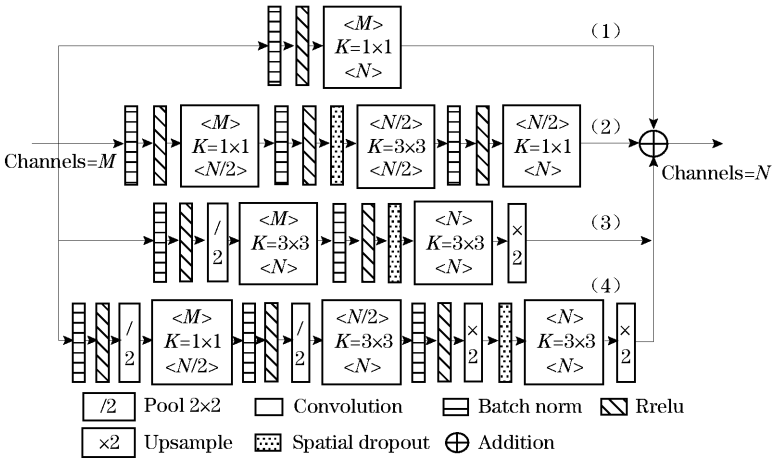

Fig. 5 Schematic diagram of MSRM

ches after being upsampled to the size of the input feature map, and then sent to the next module. Branch(3) reduces the feature map to one-half of the input feature map. Branches(1) and (2) do not change the size of the feature map. Therefore, the learning of images at different scales is achieved.

2.4 Multi-residual module stacked hourglass network

2.4.1 Multi-residual module hourglass sub-network

In the previous method, the hourglass sub-network (shown in Fig. 1) only use one residual module, so it can only capture a single scale visual mode or semantic information, and the effective field of the network is small.

In this paper, three different functional residual modules are used as the building blocks of the hourglass sub-network. Our hourglass sub-network is named as multi-residual module hourglass sub-network, the schematic diagram shown in Fig. 6 which can better capture the multi-scale visual mode and richer feature information. It consists of pool, upsample, IRM, LRFRM, and MSRM. The size of the input image of the hourglass sub-network is 64×64. The LRFRM and the MSRM are used to capture information on a lager range and on various scales, while the IRM is only used when the resolution is reduced to a minimum.

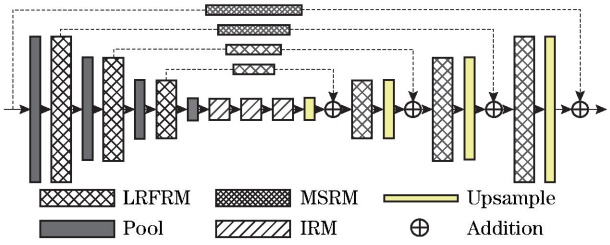

Fig. 6 Schematic diagram of multi-residual module hourglass sub-network

2.4.2 Multi-residual module stacked hourglass network

Fig. 7 shows the overall framework of our MRSH proposed in this paper, which is composed of two parts. The first half of the network uses convolution, LRFRM, and pool to capture the information on the feature map and learn the human pose estimation. The second half is composed of multiple hourglass sub-networks, and intermediate supervision is performed at the end of each hourglass. This repeated high-low, low-high processing creates a stacked hourglass network, that is, our multi-residual module stacked hourglass network. The input picture size of the network is 256×256; after a convolution layer (the size of the convolution kernel is 7×7, the step length is 2, the feature map size becomes 128×128. Then we use LRFRM to obtain the feature information, and the feature map size is changed to 64×64 after pool. Finally, the feature map passes through

two LRFRMs and enter the stacked hourglass sub-network for further feature learning. The number of hourglass sub-networks finally used in this paper is 4, achieving high precision on two common datasets.

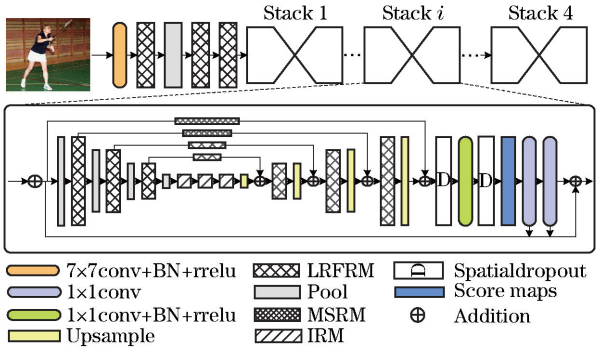

Fig. 7 Schematic diagram of MRSH

3 Joint Prediction

We use score maps to regress the predicted position of the joint points. Each hourglass sub-network produces an estimated score map corresponding to K body joint points. We use $v = \{v_k\}_{k=1}^{K}$ to represent the true label position of a joint point, in which $v_k = (x_k, y_k)$ represents the true label position of the k-th joint point in the image, so the score map of the real position is obtained by placing the Gaussian peak at the true position of the body k-th joint point. The calculation formula is

$$S_k(t) \sim N(v_k, \Sigma) \tag{4}$$

where $t \in R^2$ denotes the location and Σ is the unit matrix. For K joint points, each hourglass sub-network predicts K score maps, that is, $\hat{S} = \{\hat{S}_k\}_{k=1}^{K}$. Therefore the loss function at the end of each hourglass is defined as

$$L = \sum_{m=1}^{M} \sum_{k=1}^{K} \| S_k - \hat{S}_k \|_2^2 \tag{5}$$

where M is the number of samples. During inference, we obtain the predicted joint position $\hat{v}_k$ from the predicted score maps generated by the last stack of hourglasses by obtaining the position with the largest score as

$$\hat{v}_k = \arg\max_t \hat{S}_k(t), \ k = 1, 2, \cdots, K \tag{6}$$

4 Datasets and Criteria

4.1 Datasets

We tested the performance of network models on two widely used public datasets: Leeds Sports Poses (LSP) and MPII Human Pose.

① LSP datasets: We represent the LSP dataset[11] and its extended dataset[12] as LSP datasets. The LSP dataset is a sample of images collected from Flickr, each image annotating 14 whole body joints and containing universally recognized poses that are difficult to estimate. The data set consists of 11 000 training image samples and 1 000 test image samples, and the experiment uses person-centered (PC) annotations for training. Adding MPII data significantly improves our performance, due to its labeling quality being much better than LSP. Because of the noisy labels in the LSP dataset, we use the LSP dataset to train with the MPII Human Pose dataset and test on the LSP test set.

② MPII datasets: The MPII Human Pose dataset[13] includes approximately 25 000 images, including over 40 000 annotated poses. These images were collected from YouTube videos, each of which annotated 16 whole body joints, covering daily human activities with highly articulated human poses. Since the test set of the MPII dataset is not disclosed, the test results for the MPII dataset are all from the official website.

4.2 Criteria

Two criteria are used in the experiment to evaluate the performance of the human pose estimation method: percentage of corrected parts (PCP)[14-15], percentage of corrected keypoints (PCK)[14].

① PCP: A widely-used criterion for human pose estimation which evaluates the position accuracy of body parts. If the predicted part end points are within half of the part length from the ground truth part end points, then the body part is determined to be correct.

② PCK: Although PCP is the preferred criterion for pose evaluation, it has the drawback of penalizing shorter limbs. Therefore, PCK is introduced to measure the detection rate of body joints. PCK evaluates the position accuracy of the human joints, where a joint is considered as detected if the distance between the detected joint and the real joint is less than a fraction of the torso diameter. The torso diameter is defined as the distance between the left shoulder and the right hip of each real pose.

In our experiments, we follow the official benchmark evaluation protocols. The official benchmark on the MPII dataset adopts PCKh (using portion of head length as reference) at 0.5, while the official benchmark on the LSP dataset adopts PCP at 0.5 and PCK at 0.2.

5 Experiments and Analysis

To reduce over-fitting, we augment the training data by rotating, scaling and flipping, so that the performance of the model is more robust. All the models are trained using Torch. The initial learning rate is 2.5×10^{-4} and the parameters are optimized by the RMSProp algorithm. We use a GeForce GTX Titan X GPU to train the model on the LSP dataset and the MPII dataset.

5.1 Performance analysis of MRSH

In order to evaluate the influence of the hourglass sub-network in MRSH on human pose estimation, we use different numbers of hourglass sub-networks on the LSP dataset under the same conditions. Tab. 1 and Tab. 2 respectively show the experimental results of the network training model composed of different numbers of hourglass sub-networks under the two evaluation criteria of PCP and PCK. Fig. 8 and Fig. 9 are PCK curves and PCP curves, where the abscissa indicates different thresholds, and the ordinate indicates the accuracy of each joint.

Tab. 1 PCK results on LSP data

Number of sub-hourglass	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCK
2	98.3	92.7	88.7	85.8	92.3	93.3	92.5	91.9
4	98.3	93.7	90.0	88.0	93.6	94.3	93.2	93.0

Tab. 2 PCP results on LSP data

Number of sub-hourglass	Torso	Upper leg	Lower leg	Upper arm	Lower arm	Head	PCP
2	98.8	94.8	92.4	87.7	80.3	96.4	90.5
4	98.9	95.5	93.6	89.0	82.5	96.8	91.7

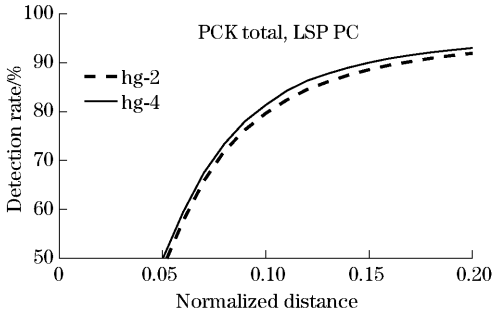

Fig. 8 PCK curve

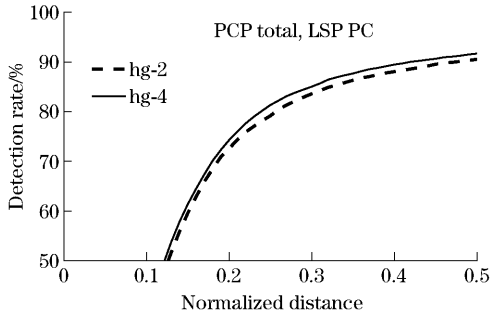

Fig. 9 PCP curve

5.2 Comparative experiment

① Comparative experiment on LSP

Our method is compared with other pose estimation methods in recent years. The comparative experiment results under the two criteria of PCK and PCP on the LSP data set are given in Tab. 3 and Tab. 4. Fig. 10 shows the PCK plots for

all methods in Tab. 3, and Fig. 11 shows the PCP plots for all of the methods in Tab. 4.

② Comparative experiment on MPII

In the experiment, several pose estimation methods related to the MPII dataset in recent years were selected and compared with our method. The comparison results under the PCKh standard are given in Tab. 5.

Tab. 3 PCK comparison results on LSP data

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCK
Ref. [16]	92.4	75.3	65.3	64.0	75.7	68.3	70.4	73.0
Ref. [17]	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6
Ref. [18]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Ref. [19]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Ref. [20]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Ref. [21]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Ref. [9]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Ref. [3]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Ref. [22]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Ours	98.3	93.7	90.0	88.0	93.6	94.3	93.2	93.0

Tab. 4 PCP comparison results on LSP data

Method	Torso	Upper leg	Lower leg	Upper arm	Lower arm	Head	PCP
Ref. [16]	95.4	77.7	69.8	62.8	49.1	86.6	70.1
Ref. [17]	95.6	78.5	71.8	72.2	61.8	83.9	74.8
Ref. [19]	96.0	86.7	82.2	79.4	69.4	89.4	82.1
Ref. [20]	97.0	88.8	82.0	82.4	71.8	95.8	84.3
Ref. [18]	98.0	93.1	88.1	82.9	72.6	83.0	85.4
Ref. [21]	97.0	90.6	86.9	86.1	79.5	95.4	87.8
Ref. [9]	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Ref. [3]	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Ref. [22]	98.4	95.0	92.8	88.5	81.2	95.7	90.9
Ours	98.9	95.5	93.6	89.0	82.5	96.8	91.7

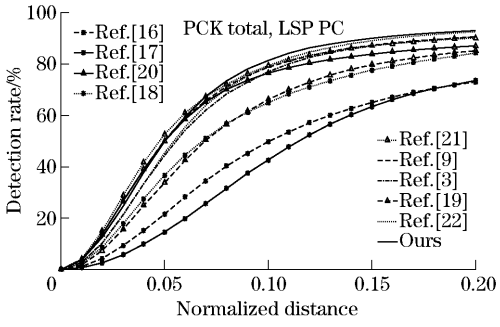

Fig. 10 PCK curves on LSP data

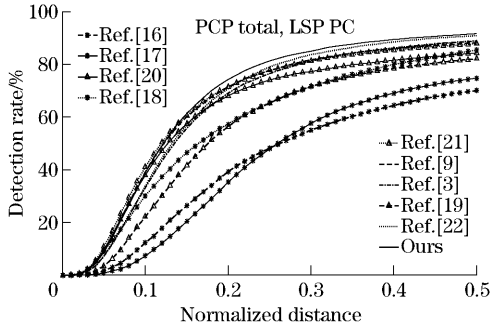

Fig. 11 PCP curves on LSP data

The results in Tab. 5 are all from the MPII official website. Ref. [10] and Ref. [22] also use the hourglass network but the final result is 8 stacked hourglasses, and the final result of this paper is 4 stacked hourglasses.

Tab. 5 PCKh comparison results on MPII data

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCKh
Ref. [5]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Ref. [23]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Ref. [20]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Ref. [24]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Ref. [25]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Ref. [26]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Ref. [19]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Ref. [21]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Ref. [9]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Ref. [3]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Ref. [10]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ref. [27]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Ref. [22]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Ours	97.9	96.1	90.8	86.0	89.9	86.3	82.8	90.4

5.3 Analysis of results

According to the results in Tab. 1 and Tab. 2, it can be seen that an increase in the number of hourglass sub-networks can effectively improve the performance of the network. This is because the characteristics learned by a single hourglass sub-network are limited and the network's receptive field is small, while the stacked hourglass network can learn more feature information. From the results in Tab. 3, Tab. 4, and Tab. 5, we can know that the method in our paper is competitive. We can achieve such good results because our MRSH can fully learn the feature information of the image at various scales during the training process, and the correlation between the joint points is learned in a large enough receptive field, which greatly improves the accuracy of the human body posture estimation.

6 Conclusion

This paper proposes a multi-residual module stacked hourglass network for human pose estimation. The network uses three different residual modules to effectively capture image feature information and visual information of images at various scales and more accurately predict joint coordinates. Our approach was tested on two challenging datasets and achieved the highest accuracy under both widely used criteria. However, some failures were observed in the experiment, including multiple people or rare poses in the picture. Handling multiple people in a single end-to-end architecture is also a challenging issue, and our future work will explore the multitude of images and the estimation of human poses for joint occlusion.

References:

[1] Tompson J, Jain A, Lecun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]//28th Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2014.

[2] Chu X, Ouyang W, Li H, et al. Structured feature learning for pose estimation[C]//29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[3] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[4] Pfister T, Charles J, Zisserman A. Flowing ConvNets for human pose estimation in videos[C]//15th IEEE International Conference on Computer Vision, Santiago, Chile, 2016.

[5] Carreira J, Agrawal P, Fragkiadaki K, et al. Human pose estimation with iterative error feedback[C]//29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//29th IEEE Conference

on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[7] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[8] Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015.

[9] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[10] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[11] Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation [C]// IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington D C, USA, 2011.

[12] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation[C]//21st British Machine Vision Conference, Aberystwyth, UK, 2010.

[13] Andriluka M, Pishchulin L, Gehler P, et al. 2D human pose estimation: New benchmark and state of the art analysis[C]// IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014.

[14] Yang Y, Ramanan D. Articulated human detection with flexible mixtures of parts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12):2878 –2890.

[15] Eichner M, Marin-jimenez M, Zisserman A, et al. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images[J]. International Journal of Computer Vision, 2012, 99(2): 190 –214.

[16] Fan X C, Zheng K, Lin Y W, et al. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015.

[17] Yang W, Ouyang W L, Li H S, et al. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[18] Yu X, Zhou F, Chandraker M. Deep deformation network for object landmark localization[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France, 2016.

[19] Belagiannis V, Zisserman A. Recurrent human pose estimation[C]// Proceedings of the International Conference on Automatic Face and Gesture Recognition(FG), Washington D C, USA, 2017.

[20] Pishchulin L, Insafutdinov E, Tang S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[21] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[22] Chu X, Yang W, Ouyang W, et al. Multi-context attention for human pose estimation[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.

[23] Hu P, Ramanan D. Bottom-up and top-down reasoning with hierarchical rectified gaussians[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016.

[24] Lifshitz I, Fetaya E, Ullman S. Human pose estimation using deep consensus voting[C]//14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[25] Gkioxari G, Toshev A, Jaitly N. Chained predictions using convolutional neural networks[C]// 14th European Conference on Computer Vision, Springer Verlag, Cham, France,2016.

[26] Rafi U, Leibe B, Gall J, et al. An efficient convolutional network for human pose estimation[C]// 27th British Machine Vision Conference, York, UK, 2016.

[27] Ning G, Zhang Z, He Z. Knowledge-guided deep fractal neural networks for human pose estimation [J]. IEEE Transactions on Multimedia, 2017, 20 (5):1246 –1259.

(**Edited by** Jianying Cai)