Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 1, 123–133
DOI: 10.34768/amcs-2021-0009

A SINGLE UPPER LIMB POSE ESTIMATION METHOD BASED ON THE IMPROVED STACKED HOURGLASS NETWORK

GANG PENG [a,b], YUEZHI ZHENG [a,b], JIANFENG LI [a,b,*], JIN YANG [a,b]

[a]Key Laboratory of Image Processing and Intelligent Control
Ministry of Education
Wuhan 430074, China

[b]School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
No. 1037 Luoyu Road, Hongshan District, Wuhan 430074, China
e-mail: {m201972557,penggang}@hust.edu.cn

At present, most high-accuracy single-person pose estimation methods have high computational complexity and insufficient real-time performance due to the complex structure of the network model. However, a single-person pose estimation method with high real-time performance also needs to improve its accuracy due to the simple structure of the network model. It is currently difficult to achieve both high accuracy and real-time performance in single-person pose estimation. For use in human–machine cooperative operations, this paper proposes a single-person upper limb pose estimation method based on an end-to-end approach for accurate and real-time limb pose estimation. Using the stacked hourglass network model, a single-person upper limb skeleton key point detection model is designed. A deconvolution layer is employed to replace the up-sampling operation of the hourglass module in the original model, solving the problem of rough feature maps. Integral regression is used to calculate the position coordinates of key points of the skeleton, reducing quantization errors and calculations. Experiments show that the developed single-person upper limb skeleton key point detection model achieves high accuracy and that the pose estimation method based on the end-to-end approach provides high accuracy and real-time performance.

Keywords: convolutional neural network, stacked hourglass network, skeleton key point, single upper limb pose estimation, human–machine coordination.

1. Introduction

In human–machine cooperative operations, the robot must estimate the pose of a single upper limb of the operator accurately and in real time and provide the pose information for trajectory prediction (Hu *et al.*, 2019; Zhou *et al.*, 2020) of the upper limb, to enable safe human–machine cooperation without collision (Zlatanski *et al.*, 2019).

In general, when using a convolutional neural network (Li *et al.*, 2017; Ning *et al.*, 2020) for pose estimation, the position coordinates of the key points of the human skeleton are directly regressed using the input image or video. Toshev and Szegedy (2015) proposed a human pose estimation method based on the AlexNet network framework, which represents the human pose estimation problem as the regression of key points of the human skeleton. Subsequently, Fan *et al.* (2015) proposed a dual-source deep convolution neural network in which the local parts were combined with the overall view, enabling more accurate human pose estimation. However, the direct regression method is not suitable for low-resolution images, has high computational complexity, and has difficulty ensuring the accuracy of the position coordinates of the key points. Therefore, based on the heat map method (Tompson *et al.*, 2015; Hu and Ramanan, 2015; Lifshitz *et al.*, 2016), Pfister *et al.* (2015) proposed a deeper convolutional neural network for human pose estimation. This method successfully transforms the problem of human pose estimation into one of human skeleton key point detection.

[*] Corresponding author

Later, Yang *et al*. (2016) proposed an end-to-end human pose estimation framework, in which a deep convolution neural network was combined with a tree structure diagram model, but the calculation efficiency of the model was low and the real-time performance required improvement. Further, Wei *et al*. (2016) proposed a convolutional pose machine model based on a convolutional neural network to remedy the low efficiency of the graph model and to make reasonable use of the spatial position information, texture information, and intermediate constraint relationship of the human body structure. The model abandons the graph method, uses a large convolution kernel to enhance the receptive field, and uses multi-stage regression to improve the accuracy of pose estimation. However, the large convolution kernel causes high computer resource consumption. To overcome this issue, Newell *et al*. (2016) proposed a multi-stage regression stacked hourglass network model in which the multi-scale feature method was used to capture the spatial position information of each key point of the human skeleton, yielding the position coordinates of each key point. This method greatly improved the receptive field and reduced the amount of calculation.

Subsequently, Chu *et al*. (2017) designed a novel hourglass residual unit, in which the stacked hourglass network model and attention mechanism were combined to solve the problem of incorrect estimation under a complex background or self-occlusion. Simultaneously, Yang *et al*. (2017) used the pyramid residual module based on the stacked hourglass network model and studied the multi-branch network weight initialization method to enhance the accuracy of human skeleton key point detection. Xiao *et al*. (2018) designed a simple baseline for easy and efficient human posture estimation in top-down mode. Sun *et al*. (2019) designed a neural network called HRNet, which has a unique parallel structure and can maintain a high-resolution representation at all times, significantly improving the effectiveness of pose recognition. Its computation is related to the network size. Zhang *et al*. (2019) proposed a model training method of fast pose distillation (FPD), which can train ultrasmall human pose neural networks more effectively and maintain sufficient accuracy, but its use of other networks to correct manual labeling is not reasonable enough. Artacho and Savakis (2020) presented the UniPose and UniPose-lstm architectures for single-image and video pose estimation. UniPose uses WASP to improve the accuracy of subject pose estimation.

To achieve the accuracy and real-time performance requirements of single upper limb pose estimation in human–robot collaboration, this paper proposes a single upper limb pose estimation method based on an end-to-end approach. A single-person upper limb skeleton key point detection model was designed using a stacked hourglass network model with high accuracy and real-time performance, and the hourglass module and human skeleton key point coordinate calculation method in the detection model were improved to increase the detection accuracy. A deconvolution layer (Long *et al*., 2015) was employed to replace the up-sampling operation of the hourglass module in the original model, solving the problem of rough feature maps. Integral regression (Sun *et al*., 2018) was used to calculate the position coordinates of key points of the skeleton, consequently reducing quantization errors and calculations. Experiments showed that the improved single upper limb skeleton key point detection model is effective and that the single upper limb pose estimation method based on the end-to-end approach provides high accuracy and real-time performance.

The specific content and structure of this paper are as follows. Section 1 introduces the background and significance of the study, expounds the research status of single-person pose estimation, and analyzes the key technical issues involved in this paper. Section 2 describes the fundamentals of human posture estimation. Section 3 introduces two methods for estimating the pose of a single person upper limb. Section 4 describes the design and improvement of a single upper limb skeleton key point detection model. Section 5 verifies the improved detection model described in Section 4, conducts experiments comparing the end-to-end approach with the cascade approach, and conducts comparative experiments on two methods of single-person upper limb pose estimation, and analyzes the experimental results. Section 6 summarizes the main work of this paper, and analyzes the limitations and expectations of the study.

2. Fundamentals of human posture estimation

2.1. Structural model of the human posture. There are four common models for human postural structures: stick models, carton models, geometric models, and 3D fine models, as shown in Fig. 1.

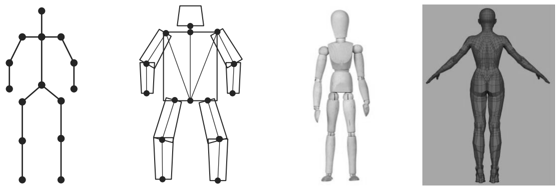


Fig. 1. Common structural models of the human body.

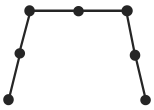


Fig. 2. Single upper limb posture structural model.

Table 1. Common datasets for single-person pose estimation.

Dataset	Basic characteristics	Information and number of key markers
FLIC	Number of samples: 5K, whole body	Coordinates, visibility, number: 9
LSP	Number of samples: 2K, whole body	Coordinates, visibility, number: 14
MPII	Number of samples: 25K, whole body	Coordinates, visibility, number: 16
COCO	Number of samples: $\geq$30W, whole body	Coordinates, visibility, number: 17

Fig. 3. Some samples from the image library.

For the single upper limb posture estimation method used in this study, the aim is to detect seven skeletal key points of the human upper limb and connect the key points into a single upper limb structure. The seven skeletal key points include right and left wrist key points, right and left elbow key points, right and left shoulder key points and neck key points. In this study, the stick model (Andriluka *et al.*, 2014) was used to characterize the human posture, and the structural model of the single upper limb posture is shown in Fig. 2.

2.2. Production of data sets. Among the available datasets for single-person pose estimation, commonly used datasets are FLIC, LSP, and MPII (Andriluka *et al.*, 2014), as shown above in Table 1. Since the datasets mentioned above are open source, deep learning can rely on these powerful datasets to improve the performance of human pose estimation. Before using the MPII dataset, the dataset needs to be pre-processed. The annotation information related to this experiment is written into a JSON format annotation file. The annotation file contains image information, body position information, body head position information, data information on the key points of the upper limb skeleton of the human body, and training set or test set annotation data. Among them, the data information of the skeletal key point of the upper limb contains the ID number, position coordinates and visibility of the skeletal key point.

Apart from the MPII dataset, a self-made dataset is also necessary. By simulating an indoor environment in which robots and humans operate together, a monocular camera is used to capture images of single human upper limbs and create a library of images of single human upper limbs. A total of 2,500 single upper limb images were included in the image library, and the images were numbered IMG_0000-2499. Some sample images in the image library are shown in Fig. 3. In the established image library, the Labelimg image annotation software was used to annotate the position of the human body, the position of the human head, and the position of key points of the upper limb skeleton. For the allocation of the training and test sets, a ratio of 4 to 1 was used to assign the training set and test set, and the relevant information obtained above was written to the annotation file.

2.3. Evaluation criteria for single upper limb posture estimation. In this study, the percentage of correct keypoints head length (PCKh) metric (Andriluka *et al.*, 2014) was used as an evaluation criterion for single upper limb posture estimation. PCKh specifically means the percentage of detections that fall within a normalized distance of the ground truth. This can be calculated in the following manner:

$$\begin{aligned}&\mathrm{PCKh@}\sigma\\&=\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}1\left(\frac{\parallel y_k^i-\hat{y}_k^i\parallel^2}{\parallel y_{\mathrm{lhip}}^i-y_{\mathrm{rsho}}^i\parallel^2}\le\sigma\right)\right).\qquad(1)\end{aligned}$$

Here, N is the number of samples, k is the k-th skeletal key, $\parallel y_k^i-\hat{y}_k^i\parallel^2$ is the distance between the predicted position coordinates of the skeletal key and the true position coordinates, and $\parallel y_{\mathrm{lhip}}^i-y_{\mathrm{rsho}}^i\parallel^2$ is the longest distance of the human head; σ represents a threshold value; in general, σ is 0.5.

3. Single upper limb pose estimation

3.1. Single upper limb pose estimation method based on the end-to-end approach. The single upper limb pose estimation method based on the end-to-end approach uses a single upper limb skeleton key point detection model. The process proceeds as follows:

(i) Compare the numbers of rows and columns in the input image to obtain a larger value of M. Then fill the input image with M rows and M columns and adjust the filled square image to 256×256 pixels. Here, the input size of the detection model is 256×256.

(ii) Using the image processed in the first step as the input of the single upper limb skeleton key point detection model, detect the key points and obtain the position coordinates of the key points.

(iii) Connect the single upper limb skeleton key points obtained in the second step in a single upper limb pose structure model based on the position relationship between the skeletal key points and the connection between them in Fig. 1.

The flow chart of the method based on the end-to-end approach is shown in Fig. 4.

3.2. Single upper limb pose estimation method based on the cascade method. The single upper limb pose estimation based on the cascade method uses a human detector and a single upper limb skeleton key point detection model. The specific process is as follows: Use the YOLOv3 network model (Redmon and Farhadi, 2018) as a human body detector to detect a single human body and to obtain a human body detection frame; after expanding the human body detection frame by 15%, cut the input image according to the expanded human body detection frame to obtain the body image. The remaining steps are the same as those of the end-to-end approach.

The flow chart of the method based on the cascade method is shown in Fig. 5.

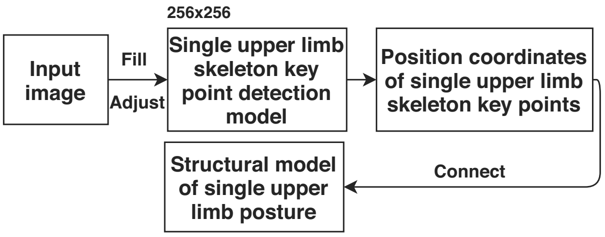

256x256
Input image
Fill
Adjust
Single upper limb skeleton key point detection model
Position coordinates of single upper limb skeleton key points
Connect
Structural model of single upper limb posture

Fig. 4. Flow chart of single-person upper limb pose estimation method based on the end-to-end approach.

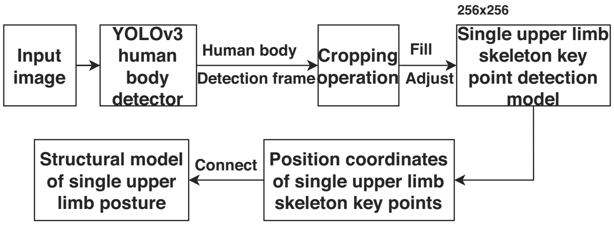

256x256
Input image
YOLOv3 human body detector
Human body
Detection frame
Cropping operation
Fill
Adjust
Single upper limb skeleton key point detection model
Position coordinates of single upper limb skeleton key points
Connect
Structural model of single upper limb posture

Fig. 5. Flow chart of single upper limb pose estimation method based on the cascade method.

4. Single upper limb skeleton key point detection

The following introduces the method of designing the single upper limb skeleton key point detection model.

4.1. Original design of the detection model. The process of designing the single upper limb skeleton key point detection model is as follows.

4.1.1. Structural design of the model. This study was based on the model structure of a stacked hourglass network, and a single upper limb skeleton key point detection model was designed that outputs only seven key points of the human upper limb. These points include the left and right wrists, the left and right elbows, the left and right shoulders, and neck key points. To verify that the stacked hourglass network model composed of eight first-order hourglass modules provides better accuracy and real-time detection performance for the skeleton key points, a number of single upper limb skeleton key point detection models were designed with different numbers and orders of hourglass modules and relevant experiments were performed. Table 2 shows the accuracy and real-time performances of the detection models with different hourglass module numbers and orders. The models are named using the form sh_ [number of hourglass modules] [order of hourglass modules], and each model was trained on the MPII data set and fine-tuned using the self-made training set. The times in Table 2 are the processing times of the models.

Table 2. Experimental results of multiple single upper limb skeleton key point detection models.

Model	Accuracy PCKh@0.5/%			Time/
name	Shoulder	Elbow	Wrist	ms
sh21	92.9	87.5	84.4	72
sh22	93.6	88.0	84.9	109
sh24	94.3	88.7	85.6	168
sh41	94.0	88.4	84.9	117
sh42	94.5	89.0	85.6	185
sh44	95.2	89.6	86.4	293
sh81	94.7	89.1	85.8	175

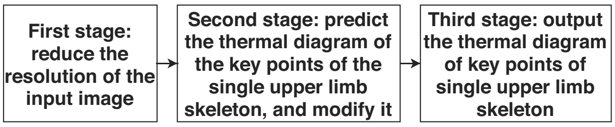

First stage: reduce the resolution of the input image
Second stage: predict the thermal diagram of the key points of the single upper limb skeleton, and modify it
Third stage: output the thermal diagram of key points of single upper limb skeleton

Fig. 6. Structural flow chart of the single upper limb skeleton key point detection model.

It can be seen from Table 2 that as the number and order of hourglass modules increase, the accuracy of the single upper limb skeleton key point detection model increases, but the real-time performance requires improvement. The single upper limb skeleton key point detection model composed of eight one-stage hourglass modules provides reasonable accuracy and real-time performance. Therefore, in this study, eight first-order hourglass modules were used to design the detection model according to the structure of a stacked hourglass network.

It can be seen from Fig. 6 that the structural flow chart of the single upper limb skeleton key point detection model is divided into three stages. The structure of this detection model is shown schematically in Fig. 7. The first stage (Fig. 7(a)) is mainly used to convolve the input image, which is then passed through the residual module and lower sampling layer, to reduce the resolution of the input image. The main purpose of the second stage (Fig. 7(b)) is to stack seven modules, including a first-order hourglass module, a residual module, a convolution layer, a batch normalization module, and an activation function layer, to predict the heat map of the key points of the single upper limb skeleton and to revise the heat map continuously. The third stage (Fig. 7(c)) is mainly composed of a first-order hourglass module, a residual module, a convolution layer, a batch normalization module, and an activation function layer. The purpose is to output the heat map of the key points of the single upper limb skeleton after constant correction.

The structure diagrams of the residual and first-order hourglass modules are shown in Fig. 8.

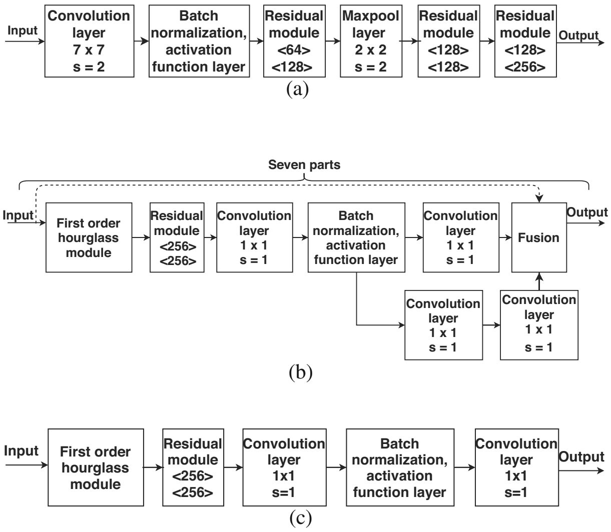

Input
Convolution layer 7 x 7 s = 2
Batch normalization, activation function layer
Residual module <64> <128>
Maxpool layer 2 x 2 s = 2
Residual module <128> <128>
Residual module <128> <256>
Output
(a)
Seven parts
Input
First order hourglass module
Residual module <256> <256>
Convolution layer 1 x 1 s = 1
Batch normalization, activation function layer
Convolution layer 1 x 1 s = 1
Fusion
Output
Convolution layer 1 x 1 s = 1
Convolution layer 1 x 1 s = 1
(b)
Input
First order hourglass module
Residual module <256> <256>
Convolution layer 1x1 s=1
Batch normalization, activation function layer
Convolution layer 1x1 s=1
Output
(c)

Fig. 7. Structural flow charts of the first (a), second (b), and third (c) stages of the single upper limb key point detection model.

4.1.2. Loss function. In the process of training the detection model of the single upper limb bone key points, the mean square error (MSE) is used as the loss function to calculate the error between the predicted and actual heat maps of the upper limb skeletal key points of the human body, so that the hourglass module can be evaluated more accurately:

$$\mathrm{MSE}^a = \frac{1}{m}\sum_{n=1}^{m}\left(\hat{y}_n^a - y_n^a\right)^2. \tag{2}$$

Here, m is the total number of pixels in the heat map of the single upper limb skeletal key points, $\hat{y_n^a}$ is the probability corresponding to each pixel position n in the predicted heat map of the a-th upper limb skeletal key points, and y_n^a is the probability corresponding to each pixel position n in the actual heat map of the a-th upper limb skeletal key points.

4.1.3. Calculation of position coordinates of the key points of a bone. In the detection model, the maximum likelihood method is used to calculate the position coordinates of the skeleton key points. The position corresponding to the pixel with the greatest probability in the heat map of the skeleton key points is taken as the position coordinate of the skeleton key point, which can be expressed as follows:

$$J_k = \arg\max_p H_k(p). \tag{3}$$

Here, J_k is the position coordinate of the k-th skeleton key point of the upper limb, p is the position in the heat map of

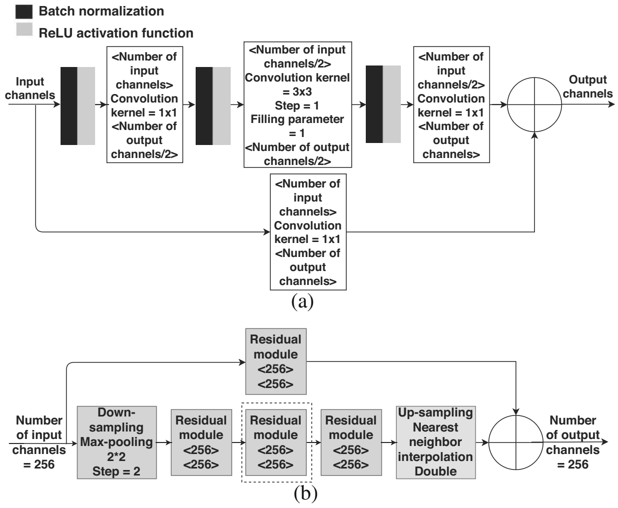

Batch normalization
ReLU activation function
Input channels
<Number of input channels> Convolution kernel = 1x1 <Number of output channels/2>
<Number of input channels/2> Convolution kernel = 3x3 Step = 1 Filling parameter = 1 <Number of output channels/2>
<Number of input channels/2> Convolution kernel = 1x1 <Number of output channels>
Output channels
<Number of input channels> Convolution kernel = 1x1 <Number of output channels>
(a)
Residual module <256> <256>
Number of input channels = 256
Down-sampling Max-pooling 2*2 Step = 2
Residual module <256> <256>
Residual module <256> <256>
Residual module <256> <256>
Up-sampling Nearest neighbor interpolation Double
Number of output channels = 256
(b)

Fig. 8. Structure diagrams of the residual (a) and first-order hourglass (b) modules.

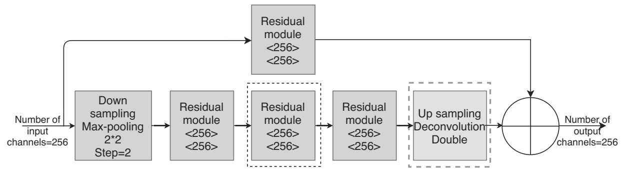

Residual module <256> <256>
Number of input channels=256
Down sampling Max-pooling 2*2 Step=2
Residual module <256> <256>
Residual module <256> <256>
Residual module <256> <256>
Up sampling Deconvolution Double
Number of output channels=256

Fig. 9. Structure diagram of the improved first-order hourglass module.

the skeleton key point, and H_k is the predicted heat map of the skeleton key point.

With the original design of the single upper limb skeleton key point detection model, the accuracies of the key point detection of the elbow and wrist of the test sample image are 89.1% and 85.8%, respectively. To increase the accuracy, it is necessary to improve the original design of the single upper limb skeleton key point detection model.

4.2. Improved detection model design. The following describes the process of improving the single-person upper limb skeleton key point detection model from two aspects.

4.2.1. Improvement of the hourglass module. The hourglass module in the original design of the detection model uses the nearest neighbor interpolation method for the up-sampling operation of the feature map. In this method, the content of the feature map is copied directly to expand the feature map, causing the feature map to be rough. Therefore, in this study, we used a deconvolution layer (Long *et al.*, 2015) to expand the feature map by slightly improving the up-sampling operation in the hourglass module to obtain a more precise feature map. The specific structure diagram is shown in Fig. 9.

The convolution kernel parameter in deconvolution layer is determined while training the network, and the relation between the dimensions of the input and output characteristic graphs of the deconvolution layer can be expressed as follows:

$$\begin{aligned}\text{Output}_{\text{size}} &= \text{stride}\cdot(\text{Input}_{\text{size}}-1)\\ &\qquad + \text{Kernel}_{\text{size}} - 2\cdot\text{padding}.\end{aligned}\tag{4}$$

Here, $\text{Output}_{\text{size}}$ is the size of the output feature map, $\text{Input}_{\text{size}}$ is the size of the input feature map, $\text{Kernel}_{\text{size}}$ is the size of the convolution kernel, and 'padding' is the fill parameter.

Finally, within the hourglass module, the feature map is up-sampled by a deconvolution layer to restore it to the same size as the input image.

4.2.2. Improvement of skeleton key point coordinate calculation. Because the original design of the detection model uses the maximum likelihood value to calculate the position coordinates of the skeleton key points, the accuracy of the detection model is easily affected by the down-sampling operation. After down-sampling, the resolution of the heat map of the key points of the single upper limb skeleton is much lower than that in the original image, leading to an irreversible quantization error.

Simultaneously, if the heat map of the key points of the human upper limb skeleton adopts a relatively high resolution, it will cause complex calculations, increased memory consumption, and low real-time performance. Therefore, in this study, integration regression (Sun *et al.*, 2018) was used instead of the maximum likelihood method to calculate the position coordinates of the key points of the skeleton, that is, to calculate the integral of all positions in the heat map of the skeleton key points of the upper limb, and the calculation results were taken as the results for the key points of the skeleton, which can be expressed as follows:

$$\begin{aligned}J_k &= \int_{p\in\Omega} p\cdot H_k'(p)\\ H_k'(p) &= \frac{e^{H_k(p)}}{\int_{q\in\Omega} e^{H_k(q)}}.\end{aligned}\tag{5}$$

In a two-dimensional heat map of the key points of the human upper limb skeleton, Eqn. (5) can be transformed into

$$J_k = \sum_{p_y=1}^{H}\sum_{p_x=1}^{W} p\cdot\frac{e^{H_k(p)}}{\int_{q\in\Omega} e^{H_k(q)}}.\tag{6}$$

Here, J_k is the position coordinate of the k-th skeleton key point of the human upper limb, p is the position of the heat map of the skeleton key point of the human upper limb, and Ω is the area of the heat map; H_k' is the normalized heat map.

In this study, the improved detection model was trained on the MPII data set and fine-tuned on the self-made training set. In the self-made test set, the key points of the single upper limb skeleton in the test sample image were detected, and the detection effects are shown in Fig. 10.

5. Experimental results and analysis

5.1. Setting of experimental parameters and hardware conditions.

5.1.1. Setting of experimental parameters. To train and fine-tune the original and improved single upper limb skeleton key point detection models, the training parameters were set as shown in Table 3.

5.1.2. Hardware conditions in the experiment. The hardware conditions used in the comparison experiments in this study are shown in Table 4.

5.2. Comparison experiments and results analysis.

The comparison experiments performed in this study included single upper limb skeleton key point detection model and single upper limb pose estimation method comparisons as well as the analysis of the experimental results.

5.2.1. Single upper limb skeleton key point detection model comparison. To confirm the validity of the improved model, the original model, the improved model, the cascaded pyramid network model, and the convolutional pose machine network model were compared in the same environment with the same test set. Among them, the cascaded pyramid network model and the convolution attitude network model were retrained on the MPII dataset and a self-made training set. The base network for the cascading pyramid network model uses ResNet-101 with an input image size of 384*288, whereas the convolutional pose machine network model is the generic model. In the contrast experiment, the test sample image was zoomed to the appropriate size and then directly input into the above four models. The accuracy curves of four models used to detect the key points of the elbow and wrist in the test sample image are shown in Fig. 11.

Using 0.5 as the threshold value of PCKh, the accuracies and real-time performances of four single upper limb skeleton key point detection models were obtained using test samples and are presented in Table 5. The times in the table are the model processing times.

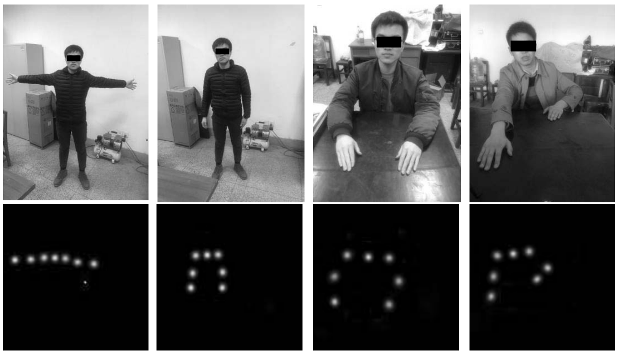


Fig. 10. Detection effects of the single upper limb skeleton key point detection model.

Table 3. Training parameter settings of the single upper limb skeleton key point detection model.

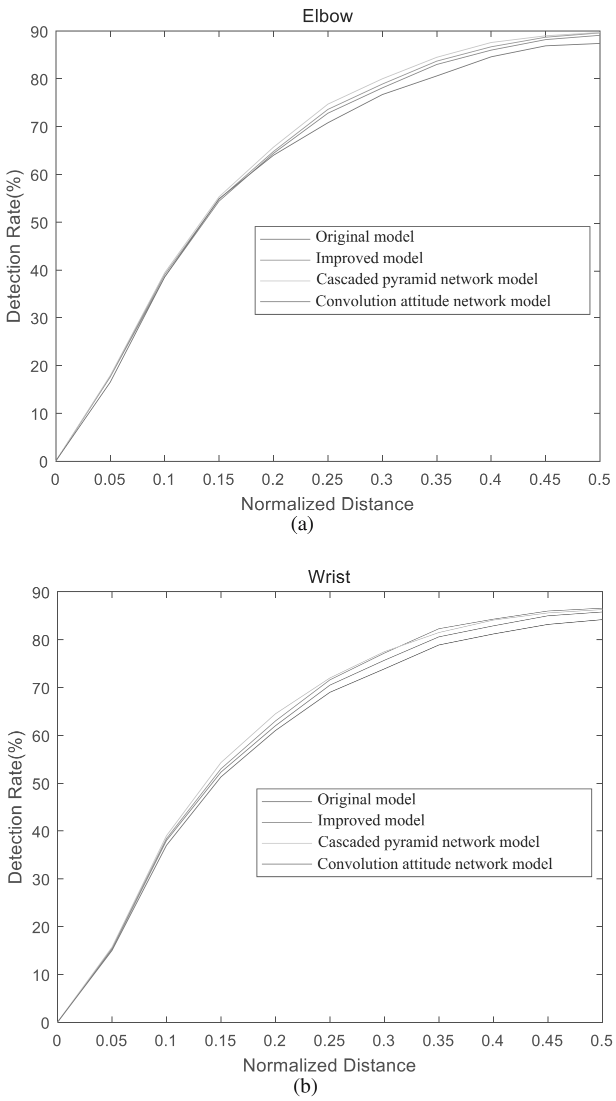
Parameter name	Set value or method
Order of hourglass module	1
Number of hourglass modules	8
Optimization method	RMS prop algorithm
Initial learning rate	0.00025
Loss function	MSE
Batch size	8
Epoch parameter	100
Number of epoch iterations	1000
Data augmentation method	Random crop, color dither, rotation

Table 4. Hardware conditions in comparison experiments.

Operating system	Ubuntu16.04 LTS
Deep learning framework	Pytorch
CPU model	i7-7700K
CPU frequency	4.2GHZ
RAM	32GB
Graphics card type	NVIDIA TITAN XP 11G
CUDA version	CUDA 8.0

As shown in Fig. 11 and Table 5, the proposed single upper limb skeleton key point detection model improves the detection accuracy. Compared with the cascaded pyramid network model, the key point detection accuracy of the improved model is only slightly different, but the real-time performance is superior. Compared with the convolutional pose machine network model, the improved model has slightly higher detection accuracy and real-time performance. Therefore, the detection model based on the improved stack hourglass network is effective, and achieves good accuracy and real-time performance.

5.2.2. Comparison of single upper limb pose estimation methods. In this study, the improved single upper limb skeleton key point detection model was



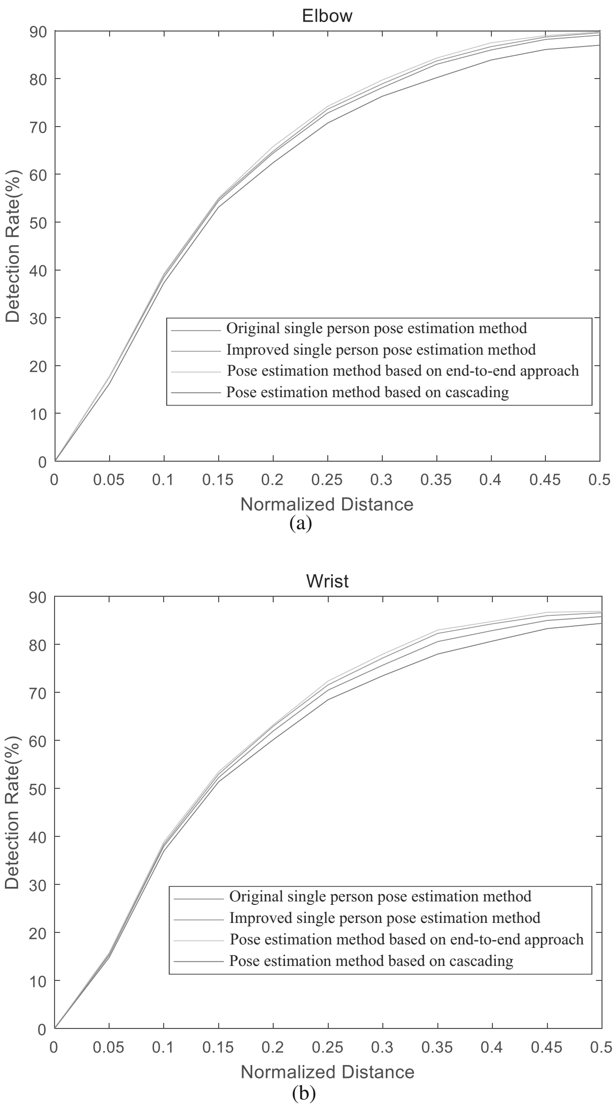
(a)

(b)

Fig. 11. Accuracy curves of four single upper limb skeleton key point detection models for the elbow (a) and wrist (b).

applied in combination with two single upper limb pose estimation methods. For the end-to-end single upper limb pose estimation method, the experiment was conducted according to the steps described in Section 3.1. For the single upper limb pose estimation method in cascade mode, the YOLOv3 network model was fine-tuned on the self-made training set, and the single upper limb pose estimation experiment was performed according to the steps described in Section 3.2. For more convincing comparisons, the original input image was zoomed to the appropriate size and then directly input into the original and improved designs of the single upper limb skeleton key point detection model.

The accuracy curves corresponding to the four methods of detecting the key points of the elbow and wrist in the test sample images are shown in Fig. 12.



(a)

(b)

Fig. 12. Detection accuracy curves of four single upper limb pose estimation methods for the elbow (a) and wrist (b).

Using 0.5 as the threshold value of PCKh, the accuracies and real-time performances of four single upper limb pose estimation methods for single upper limb skeleton key points were obtained by using test samples and are presented in Table 6.

As can be seen from Fig. 12 and Table 6, the end-to-end single upper limb pose estimation method is slightly more accurate than directly scaling the original input image to the appropriate size and inputting it into the detection model. Because the original input image is filled and then adjusted, the original horizontal-to-vertical ratio of the image can be maintained when adjusting the image, so that the adjusted image will not be deformed. This approach allows the network to extract features faster and more accurately, thus improving accuracy and real-time performance.

Table 5. Experimental results of four single upper limb skeleton key point detection models.

Model name	Accuracy PCKh@0.5/%			Time/ ms
	Shoulder	Elbow	Wrist	
Convolutional attitude machine network	94.0	87.4	84.2	398
Cascaded pyramid network	95.6	89.8	86.3	278
Original detection model	94.7	89.1	85.8	175
Improved detection model	95.4	89.6	86.6	179

Table 6. Experimental results of four single upper limb pose estimation methods.

Method name	Accuracy PCKh@0.5/%			Time/ ms
	Shoulder	Elbow	Wrist	
Original pose estimation method	94.7	89.1	85.8	175
Improved pose estimation method	95.4	89.6	86.6	179
Pose estimation method based on end-to-end approach	95.7	89.8	86.9	171
Pose estimation method based on cascading	93.2	87.0	84.4	214

Compared with the end-to-end single upper limb pose estimation method, the method obtained by cascading a human detector and an improved single upper limb skeleton key point detection model yielded poor accuracy and real-time performance. These poor results were obtained because the human detector based on the YOLOv3 network model could not accurately detect the human body in the original input image and the input image of the single upper limb skeleton key point detection model was based on the human detection results. Therefore, the human body detector incorrectly detected the human body in the original input image and used the wrong human body image as the input for the single upper limb skeleton key point detection model, reducing the detection accuracy. Furthermore, it also needs time to detect the human body in the original input image.

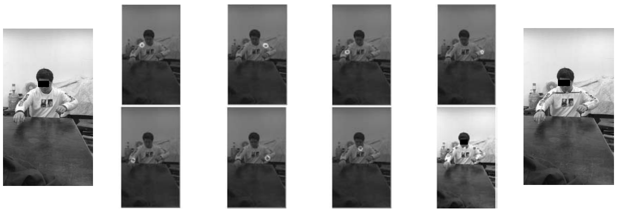


Fig. 13. Experimental results of single upper limb pose estimation based on the end-to-end approach.

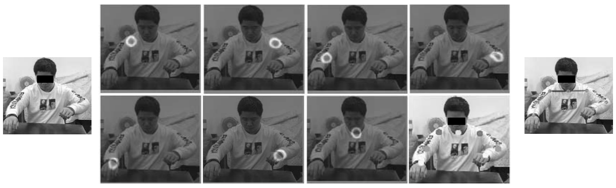


Fig. 14. Experimental results of single upper limb pose estimation based on the cascade method.

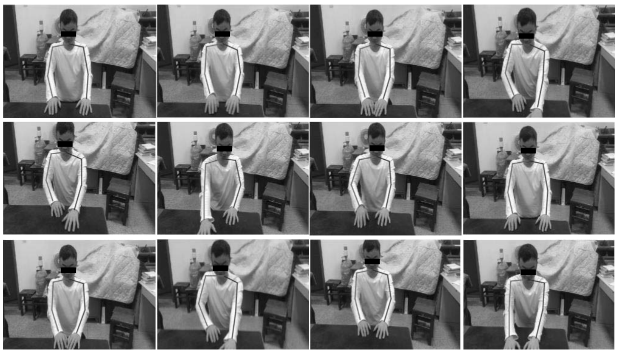


Fig. 15. Effect of single upper limb pose estimation on each frame in the same video.

In general, these results demonstrate that the single upper limb pose estimation method based on the end-to-end approach has better accuracy and real-time performance than the cascade method.

5.2.3. Experimental results of the single-person upper limb pose estimation method. The experimental results obtained using the single upper limb pose estimation methods based on the end-to-end approach and cascade method are shown in Figs. 13 and 14, respectively; the results of the pose estimation experiment for each video frame are shown in Fig. 15. It can be seen that the proposed single-person upper limb pose estimation method is feasible and effective.

6. Conclusion

Based on previous research on human–machine cooperative operation, an end-to-end single-person upper limb pose estimation method was developed in this study. Using the stacked hourglass network model, a single-person upper limb skeleton key point detection model was designed and the hourglass module and human skeleton key point coordinate calculation method were improved. Experiments confirmed the effectiveness of the improved single upper limb skeleton key point detection model. Compared with single upper limb pose estimation based on the cascade method, the proposed end-to-end single-person upper limb pose estimation method yields higher accuracy and real-time performance.

This study also has some limitations, which need to be further improved. It focuses on the estimation method of the human upper limb pose, which can provide the pose data for the subsequent trajectory prediction of upper limb movement. This will be the focus of future work.

References

Andriluka, M., Pishchulin, L., Gehler, P.V. and Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA,* pp. 3686–3693.

Artacho, B. and Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR Virtual)*, pp. 7035–7044, (online).

Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L. and Wang, X. (2017). Multi-context attention for human pose estimation, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA*, pp. 5669–5678.

Fan, X., Zheng, K., Lin, Y. and Wang, S. (2015). Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA*, pp. 1347–1355.

Hu, H., Liao, Z. and Xiao, X.C. (2019). Action recognition using multiple pooling strategies of CNN features, *Neural Processing Letters* **50**(1): 379–396.

Hu, P. and Ramanan, D. (2015). Bottom-up and top-down reasoning with convolutional latent-variable models, *ArXiv:* abs/1507.05699.

Li, C., Yung, N.H.C., Sun, X. and Lam, E.Y. (2017). Human arm pose modeling with learned features using joint convolutional neural network, *Machine Vision and Applications* **28**(1–2): 1–14.

Lifshitz, I., Fetaya, E. and Ullman, S. (2016). Human pose estimation using deep consensus voting, *European Conference on Computer Vision (ECCV), Amsterdam, Holland*, pp. 246–260.

Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA*, pp. 3431–3440.

Newell, A., Yang, K. and Deng, J. (2016). Stacked hourglass networks for human pose estimation, *European Conference on Computer Vision (ECCV), Amsterdam, Holland*, pp. 483–499.

Ning, F., Shi, Y., Cai, M. and Xu, W. (2020). Various realization methods of machine-part classification based on deep learning, *Journal of Intelligent Manufacturing* **31**(8): 2019–2032.

Pfister, T., Charles, J. and Zisserman, A. (2015). Flowing ConvNets for human pose estimation in videos, *2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pp. 1913–1921.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement, *arXiv:* 1804.02767.

Sun, K., Xiao, B., Liu, D. and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation, *Computer Vision and Pattern Recognition (CVPR), Los Angeles, USA*, pp. 5693–5703.

Sun, X., Xiao, B., Wei, F., Liang, S. and Wei, Y. (2018). Integral human pose regression, *European Conference on Computer Vision (ECCV), Munich, Germany*, pp. 529–545.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y. and Bregler, C. (2015). Efficient object localization using convolutional networks, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA*, pp. 648–656.

Toshev, A. and Szegedy, C. (2015). DeepPose: Human pose estimation via deep neural networks, *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA*, pp. 1653–1660.

Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y. (2016). Convolutional pose machines, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA*, pp. 4724–4732.

Xiao, B., Wu, H. and Wei, Y. (2018). Simple baselines for human pose estimation and tracking, *European Conference on Computer Vision (ECCV), Munich, Germany*, pp. 466–481.

Yang, W., Li, S., Ouyang, W., Li, H. and Wang, X. (2017). Learning feature pyramids for human pose estimation, *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy,* pp. 1281–1290.

Yang, W., Ouyang, W., Li, H. and Wang, X. (2016). End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA*, pp. 3073–3082.

Zhang, F., Zhu, X. and Ye, M. (2019). Fast human pose estimation, *Compter Vision and Pattern Recognition (CVPR), Los Angeles, USA*, pp. 3517–3526.

Zhou, J., Liu, J. and Zhang, M. (2020). Curve skeleton extraction via k-nearest-neighbors based contraction, *International Journal of Applied Mathematics and Computer Science* **30**(1): 123–132, DOI: 10.34768/amcs-2020-0010.

Zlatanski, M., Sommer, P., Zurfluh, F., Zadeh, S.G., Faraone, A. and Perera, N. (2019). Machine perception platform for safe human-robot collaboration, *2019 IEEE SENSORS, Montreal, Canada*, pp. 1–4.

Gang Peng (born in 1973) received his doctoral degree from the Huazhong University of Science and Technology (HUST) in 2002. Currently, he is an associate professor in the Department of Automatic Control, School of Artificial Intelligence and Automation, HUST. He is also a senior member of the China Embedded System Industry Alliance and the China Software Industry Embedded System Association, a senior member of the Chinese Electronics Association, and a member of the Intelligent Robot Professional Committee of Chinese Association for Artificial Intelligence. His research interests include intelligent robots, machine vision, multi-sensor fusion, machine learning and artificial intelligence.

Yuezhi Zheng received a Master's degree from the Huazhong University of Science and Technology, China, in 2019. His research interests include robotics and computer vision.

Jianfeng Li received a Bachelor's degree in engineering from Nanchang University, China, in 2019. He is currently a graduate student at the Department of Automatic Control, Huazhong University of Science and Technology, Wuhan, China. His research interests include robotics and computer vision.

Jin Yang received a Bachelor's degree in engineering from Tianjin Polytechnic University, China, 2019. He is currently a graduate student at the Department of Automatic Control, Huazhong University of Science and Technology, Wuhan, China. His research interests include robotic arm control and image processing.

Received: 1 April 2020
Revised: 4 August 2020
Accepted: 17 August 2020