Human Pose Estimation Based on Multi-resolution Feature Parallel Network for Public Security

Xiangru Tao
Cheng Xu
Hongzhe Liu
Zhibin Gu

Smoking detection is an essential part of safety production management. With the wide application of artificial intelligence technology in all kinds of behavior monitoring applications, the technology of real-time monitoring smoking behavior in production areas based on video is essential. In order to carry out smoking detection, it is necessary to analyze the position of key points and posture of the human body in the input image. Due to the diversity of human pose and the complex background in general scene, the accuracy of human pose estimation is not high. To predict accurate human posture information in complex backgrounds, a deep learning network is needed to obtain the feature information of different scales in the input image. The human pose estimation method based on multi-resolution feature parallel network has two parts. The first is to reduce the loss of semantic information by hole convolution and deconvolution in the part of multi-scale feature fusion. The second is to connect different resolution feature maps in the output part to generate the high-quality heat map. To solve the problem of feature loss of previous serial models, more accurate human pose estimation data can be obtained. Experiments show that the accuracy of the proposed method on the coco test set is significantly higher than that of other advanced methods. Accurate human posture estimation results can be better applied to the field of smoking detection, and the smoking behavior can be detected by artificial intelligence, and the alarm will be automatically triggered when the smoking behavior is found.

Keywords: Public security, Smoking detection, Pose estimation, Action recognition

Tob Regul Sci.™ 2021;7(5): 1049-1058
DOI: doi.org/10.18001/TRS.7.5.22

The purpose of the human body pose estimation task is to locate the key points of the human body in the image, such as wrists and ankles. Human pose estimation is also the basis of many computer vision tasks[1]. The key information obtained by human body posture estimation can be used in fields such as human-computer interaction, assisted medical treatment, motion analysis, virtual reality, and autonomous driving[2]. Human action recognition [3] can also realize the detection of dangerous behavior in production work, such as smoking detection, dangerous action detection, etc. by comparing the matching degree of human action and detection action in the video, the real-time warning can be realized. When smoking behavior occurs in monitoring, it can be identified through the deep neural network, and then alarm and prompt.

There are two main contributions to this paper. One is to achieve high-precision human pose estimation through a multi-resolution feature parallel network. The second is to classify human actions by recognizing the human posture in the image. However, in the field of public security applications, to achieve smoking detection, dangerous behavior detection, and other safety monitoring tasks.

In the following chapters, firstly, in the second section, the help of artificial intelligence detection to traditional artificial detection of smoking behavior, the research background of human posture estimation, and the current more advanced work are summarized. The main methods and the existing problems are analyzed and discussed. Then in the third section, the network structure of multi-resolution net(MRnet)

Xiangru Tao, Beijing Key Laboratory of Information Service Engineering, Beijing Union University, China. Cheng Xu, Beijing Key Laboratory of Information Service Engineering, Beijing Union University, China. Hongzhe Liu , Beijing Key Laboratory of Information Service Engineering, Beijing Union University, and Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China. Zhibin Gu, Beijing Key Laboratory of Information Service Engineering, Beijing Union University, and Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China. ŸCorresponding author: Dr Hongzhe Liu: liuhongzhe@buu.edu.cn*

and the methods of MRnet used in multi-scale feature fusion and feature graph output are proposed. At the same time, the method of action matching is introduced. In Section 4, the experiment is carried out, compared with the current advanced methods, and the experimental results are analyzed. The method of action matching experiments and the result of an example are given. Finally, in Section 5, we summarize and discuss the next research direction in the field of human pose estimation and artificial intelligence smoking detection.

RELATED WORK

For the detection of smoking behavior, it is often necessary to manually observe the screen from the monitor to judge. This method is usually inefficient, a waste of manpower, and prone to negligence. It is difficult for ordinary smoke alarm equipment to detect smoking behavior in a short time. But some professional smoking detection equipment is often expensive, the application of artificial intelligence human behavior recognition can achieve real-time detection to avoid negligence. So the vision-based human pose detection method is the key to solve the problem. With the rise of convolutional neural networks, the current mainstream human pose estimation method is based on the deep convolutional neural network[4][5][6][7][8], learning the human body feature information in the image to locate The position of the key points of the human body in the image. Compared with single-person human pose estimation, multi-person human pose estimation faces more challenges. It is necessary to calculate the number of people in the image and the corresponding position of the human body. The scale of the human body in the image is different. At the same time, the occlusion problem is also a difficult point in the human pose estimation problem. Especially in some sports and dance movements, the body may block certain branches. To get the current posture of the human body in the image and correct it, we need to pay attention to multi-scale feature fusion to improve the robustness of model prediction.

In the past, the network structure often uses a serial design. To obtain a high-resolution feature representation, low-resolution features will be up-sampled multiple times. In some networks[4], the up-sampling network will use the mirror image of the down-sampling process. version. In DeconvNet, a set of indexes is used to store the source information during convolution. In this way, information loss during up-sampling is reduced. SimpleBaseline[6] demonstrated that deconvolution can generate high-quality feature maps for heat map prediction. In some works, asymmetric up-sampling processes are also used, such as using heavy up-sampling[9] and mild down-sampling. Sampling, or light upsampling to meet the needs of different tasks. To be able to locate the key points of the human body in the image more accurately, it is also necessary to pay attention to how to ensure the quality of the high-resolution feature map in the process of extracting features from the network. The early interconnected convolutional neural network[10] contains multiple neural networks, which can divide the input into feature maps of different scales, to make full use of semantic information, but there is no information exchange between different feature maps, and no residuals are used. Connection, so the effect is often poor. Similar to the dense connection method in DenseNet[11], all layers are connected. In this way, feature fusion can be achieved. Low-resolution features can obtain high-resolution feature information, but high-resolution feature information cannot be fused. Resolution feature information.

To achieve the task of posture correction, some work will use multiple pressure sensors to detect the posture of the human body[12], and in some work[13] will use RF sensors to detect the posture of the human body, but it still needs to be on the target. Paste RFID tags to improve the robustness and accuracy of the detection results. Some jobs use Kinect sensors to recognize human posture[14], but Kinect itself contains a depth sensor, and the method cannot be used in ordinary RGB images that do not contain depth information. In the method that only uses the RGB image as input, to improve the accuracy of the human pose estimation result, it is necessary to make better use of high and low-resolution feature maps, which requires the use of multi-scale feature fusion methods. The hourglass network[4] merges the feature maps of the same resolution in the down-sampling stage in the up-sampling stage utilizing layer jump connection. The CPN network[5] gradually fuses the feature information of the low-resolution feature map into the high-resolution feature map during the up-sampling process. But the low-resolution feature map does not get the feature information of the high-resolution feature map. HRNet[7] exchanges information between feature maps of different resolutions through parallel multi-resolution subnets and repeats this process many times. HRNet can retain more feature information and make the predicted heat map results more accurate.

The human pose estimation method based on the multi-resolution feature parallel network proposed in this paper is inspired by HRNet because HRNet only retains the high-resolution features during the output process and ignores other low-resolution feature information. In response to this problem, the MRNet model

integrates several other low-resolution feature information and uses deconvolution and hole convolution in the scale fusion stage to ensure that more feature information can be retained in the multi-scale feature fusion stage. In the second stage, an action matching network is proposed, which can compare the smoking action with the monitoring action, evaluate the similarity of the two actions with a given standard, and judge the smoking behavior.

Multi-resolution Feature Parallel Network

To more accurately locate the key points of the human body in the image, provide accurate human posture data for action recognition and artificial intelligence detection. This article proposes the Multi-Resolution Net (MRNet) network. Improved the up-sampling method in the multi-scale feature fusion stage and the problem of the loss of low-resolution feature map information in the output stage. The following briefly introduces the defects of HRNet and the specific methods of improving part of MRNet.

HRNet Network Model

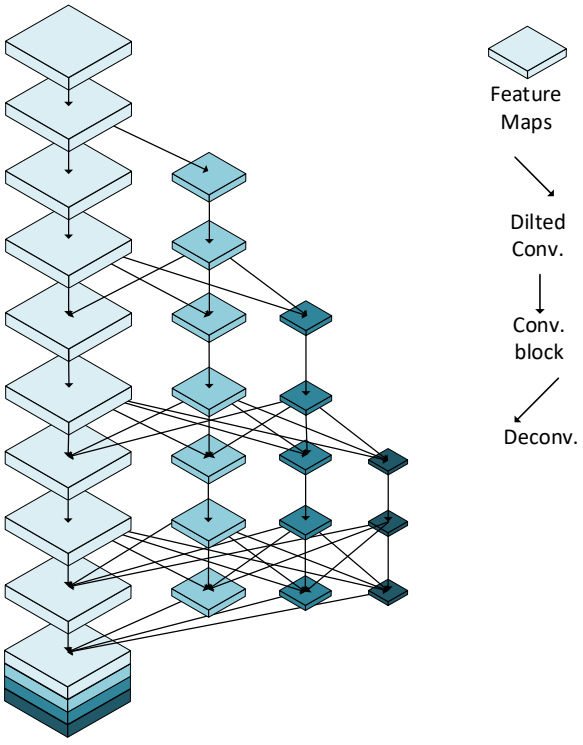
For human body pose estimation tasks, deep convolutional neural networks have shown far superior performance to other methods. Due to the requirements of human body pose estimation tasks, feature maps of different resolutions are often required, and most methods often use serial networks. Multiple up-sampling is used to extract the semantic information of feature maps of different scales, and the parallel connection of subnets of HRNet can maintain high-resolution features in the process of extracting features. However, the high-resolution prediction heat map output by HRNet still loses the low-resolution heat map information.

MRnet Network Model

The way MRNet generates the backbone network is similar to HRNet. The network is divided into five stages: backbone stage, low-resolution subnet generation stage 1, stage 2, stage 3, multi-resolution feature subnet fusion output stage.

In the main stage of processing, the input image is first sent to a convolutional network with a step size of 2, 3×3 convolution kernel for two convolutions, the size of the image is changed to 1/4 of the original size, and then the feature map is used 4 residual blocks, the residual block here uses Basicblock or Bottleblock for feature extraction, and the final output is sent to the low-resolution subnet generation stage. The low-resolution subnet generation stage is divided into three small stages. The purpose of the three stages is to generate a low-resolution subnet branch. The method used at the same time is also similar, all through 4 residual blocks (Basicblock Or Bottleblock) perform feature extraction, and then perform scale fusion on the result after feature extraction (the operation of scale fusion will be introduced in the next section). The generated subnet information will also contain the information of the different resolution subnets in the previous stage. Specifically, after stage one, two parallel subnets will be obtained. The output sizes of these two subnets are (set the original image size as [H, W]) [H/4, W/4], [H/8, W/8], a new subnet will be added after phase two, the output size of the subnet is [H/16, W/16], a new subnet will be added after phase three, the output size of the subnet It is [H/32, W/32].

Figure1. The MRNet network structure diagram shows the parallel structure of the network and the feature exchange between different size feature graphs. The size of the original graph is [H, W]. The size of the feature graphs are [H/4, W/4], [H/8, W/8], [H/16, W/16], [H/32, W/32]



After three stages of low-resolution generation, 4 subnet branches will be obtained, and the output size is[H/4,W/4],[H/8,W/8],[H/16,W/16],[H/32,W/32] feature map. In the final multi-resolution scale fusion output stage, the low-resolution subnet will be up-sampled to output a feature map of[H/4, W/4] size, and then Connect the output results of the 4 subnets, and finally generate a feature map containing all the resolution semantic information. The network structure is shown in Figure 1. Different from HRnet only using the

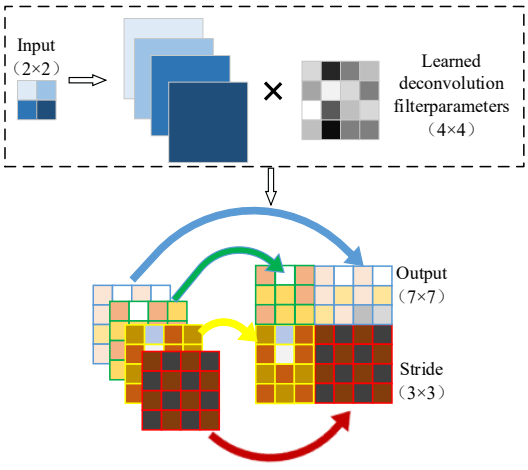
highest resolution feature map as output, the final output of MRNet contains more semantic information, which can achieve higher accuracy in pose estimation tasks.

Multi-scale Feature Fusion Method

After the high-resolution heatmaps, medium-resolution heatmaps, and low-resolution heatmaps are generated through serial subnets, information needs to be exchanged among feature maps of different scales. HRnet uses parallel multi-resolution subnets for information exchange. Up-sampling of the low-resolution feature map makes the size of the low-resolution feature map consistent with the high-resolution feature map, and the high-resolution feature map will transform the high-resolution feature map size to a low resolution through convolution operation The size of the feature map is then summed.

In the process of low-to-high feature map information exchange, HRnet uses nearest-neighbor interpolation for up-sampling. This method will affect the quality of the sampled feature map, and it is even difficult to locate small-scale targets in the original image. Inspired by this work[15], MRNet uses deconvolution to increase the scale of the feature map to make it consistent with the size of the high-resolution feature map. The operation of deconvolution is shown in Figure 2 below.

Figure 2. Deconvolution When the input image is 2×2, the output of 7×7 can be obtained when the convolution kernel is 4×4 and the step size is 3.



In this way, instead of the up-sampling method of interpolation, since the deconvolution parameters can be trained, the loss of features can be reduced during the enlargement of the small-size feature map.

In the scale fusion part of MRNet, the low-resolution feature map is transformed into the same size and number of channels as the high-resolution feature map through 1×1 convolution and deconvolution. Set the corresponding convolution kernel and step size for the 1×1 convolution and deconvolution of different resolution feature maps at different stages. Since the parameters of the deconvolution are trainable, the feature information can be better preserved when the low-resolution feature map transmits information to the high-resolution feature map. To output a feature map of a specific size, the following formula can be used for calculation:

When $(o+2p-k)\%s=0$

$$o=s(i-1)-2p+k \quad (1)$$

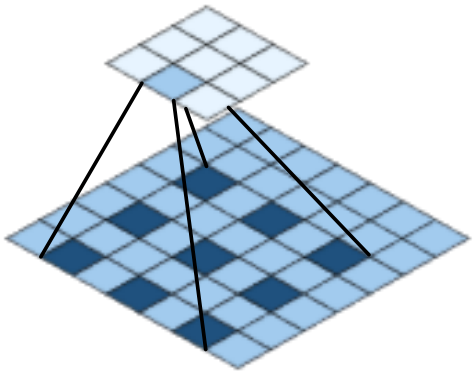
When $(o+2p-k)\%s\neq 0$

Where i is the output size, k is the size of

$$o=s(i-1)-2p+k+(o+2p-k)\%s \quad (2)$$

the convolution kernel, s is the sliding step size, p is the padding size, and o is the output size.

Figure 3. Dilated Convolution The receptive field is enlarged by convolution operation with interval



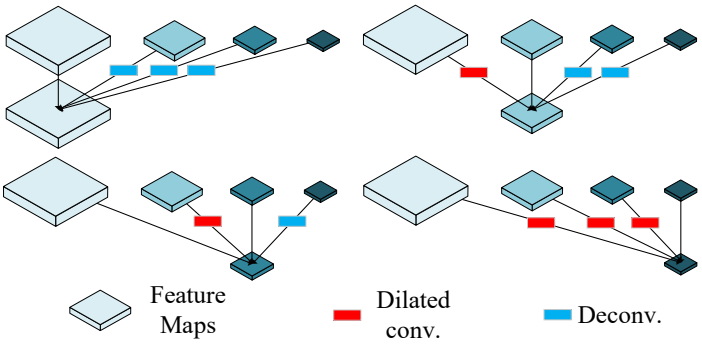
In the high-to-low feature map information exchange, using a 3×3 convolution with a step size of 2 for down-sampling can reduce the number of parameters, but it will also cause the loss of features. If a larger convolution kernel is used, although the receptive field can be increased without loss of resolution, it will greatly increase the number of parameters. To resolve this contradiction. Inspired by the method of this work[16]. We introduce dilated convolution in the process of high-to-low feature map information exchange.

Dilated convolution can be realized in two different ways. The first way is to fill 0 in the convolution kernel, and the second way is to sample the input interval. The convolution kernel is 3×3 and the interval ratio is 2 dilated convolutions. The product is shown in Figure 3. Dilated convolution can provide a larger receptive field without losing

resolution, and no additional parameters need to be introduced. In this way, context information can be obtained more effectively.

The final parallel multi-scale fusion network in MRNet is shown in Figure 4.

Figure 4. MRNet multi-scale feature fusion method Four feature fusion methods with the different resolutions are presented.

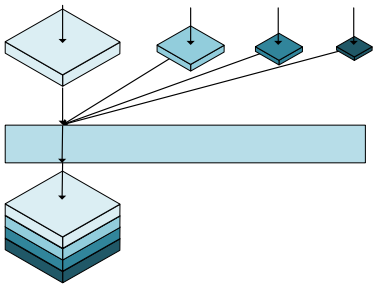


Feature Map Output Method

In the feature map output part, HRNet directly outputs the highest resolution feature representation, and the feature representations of other resolutions will be ignored. In MRNet, to retain the information of the three parallel low-resolution subnets, the three parallel low-resolution subnets are up-sampled into the size of the highest resolution image and then connected, and finally 1×1 is used. The convolution transforms the feature map into semantic segmentation categories to get the final result. As shown in Figure 5.

MRNet not only retains the high-resolution features, but also uses several other low-resolution features, and only adds a few parameters and calculations, and the resulting features will be more powerful and spatially accurate.

Figure 5. MRnet outputs the feature map. MRnet fuses the features of different resolution size feature maps with weight in the feature map output stage.



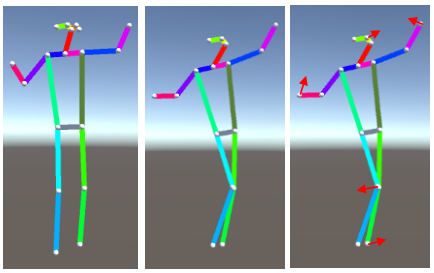
Action Recognition Method

Starting from a certain frame of the dangerous action video and the monitored person's video, the human key point information extracted from the dangerous action and the monitored person's action image is globally aligned, ignoring the error caused by the shooting position, size, and other environmental factors, and only from the difference of the two human postures to the same position for analysis. We use $Y \mid X = \{y_1 \mid x_1, y_2 \mid x_2, \cdots y_n \mid x_n\}$, It indicates the posture sequence of dangerous action and monitored action. At the same time or corrective action vector:

$$V = Y \mid X - X \qquad (3)$$

The similarity between dangerous work and the action of the monitored was judged by the guidance vector. As shown in Figure 6.

Figure 6. The schematic diagram of the guidance vector, where (a) is the dangerous action video, (b) is the action of the monitored person, and (c) the matching degree between the two actions and the guidance vector.



(a) (b) (c)

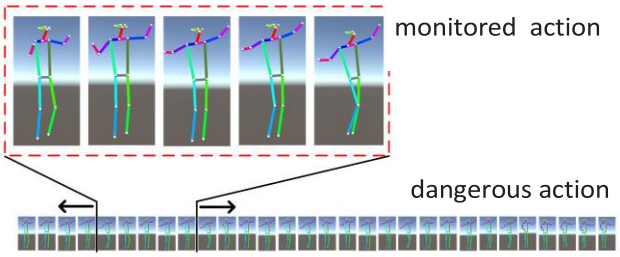
For a dangerous action in public safety, it is usually multi-frame continuous action. If we simply compare the motion differences of a single frame, we can not fully meet the requirements of motion matching. Therefore, in action matching, we should also start from the continuity of action. For the monitored person, some actions are slow, some actions are fast, which will lead to the inaccuracy of action matching results. Therefore, we need to compare

multiple action frames in the network to analyze actions. Is the time right? At 30 frames per second. In the case of time consistency, we will adopt five learner action frameworks. The errors of the key points were scored between the five action frames and the standard action. The scoring criteria are as follows:

$$S = -\sum_{i=1}^{N} \left\| (Y_i - X_i) - (\hat{Y}_i - \hat{X}_i) \right\|_2 \quad (4)$$

Among them, N represents the joint points of the corresponding human body. For example, 1 represents the left ear, 2 represents the left eye, Y_i represents the position of a joint point of dangerous action, and X_i represents the position of a joint point of the monitored person. The reciprocal of Euclidean distance between the joint position of dangerous activity and the key position of the monitored person was taken as the action standard score. Take the average of 5 points, then slide 5 frames of the monitored sequence on the dangerous action sequence, and repeat the above operation after sliding. If the score increases when sliding to the left, it is judged that the speed of the monitored action is faster than the dangerous action. If the score is increased by sliding to the right, it is judged that the speed of the monitored action is slower than the dangerous action. If the score of the guidance vector is high, the action of the monitored person can still be judged as a dangerous activity. An example diagram is shown in Figure 7.

Figure 7. The action matching diagram slides the monitored action sequence on the dangerous action sequence to match more identical actions with different speeds



EXPERIMENTAL RESULT AND ANALYSIS

Dataset: The dataset uses the COCO dataset[17]. The COCO dataset contains more than 200,000 pictures and more than 250,000 human instance objects. Each object is marked with 17 key points of the human body. Train the model on COCO train2017. There are more than 57,000 images and more than 150,000 human instance objects on COCO train2017. After the training is completed, evaluate the effectiveness of the method on COCO val2017 and COCO test-dev2017.

At the same time, the experimental data will also use the new dangerous action data set, which includes 97 groups of dangerous videos, including dangerous action videos and monitored person action videos.

Evaluation standard: The standard evaluation index is defined by the target key point similarity (OKS):

$$\frac{\sum_i \exp(-d_i^2 / 2s^2 k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (5)$$

The d_i is the Euclidean distance between the key points obtained by the ground truth and the test of the tag, v_i indicating whether the true value of the tag is visible, s representing the size of the object, and the k_i control attenuation constant of each key point. The following tests and report the standard average accuracy and recall score, AP means the average accuracy of the 10 positions at OKS=100. The AR means the average Recall score.

Training method: In the training phase, we use the same settings as HRNet to expand the human detection frame to a fixed aspect ratio, height:width=4:3, crop the human detection frame from the image, and adjust its size to 256×192 or 384×288, the data enhancement part also uses random rotation ([-45°, 45°]), random scale scaling ([0.65, 1.35]) and flipping. The optimizer chooses the Adam optimizer, the initial learning rate is 1e-3, and the learning rate is reduced to 1e-4 and 1e-5 in 170 and 200 rounds of training, respectively. A total of 210 rounds of training.

Test method: Using the top-down method in the test, first use the human body detector to detect the human object, and then predict the key points of the human body, use the same human body detector as in this method[18] for the human body detection, pass the original image and flip The average heatmaps of the image is used to calculate the final heatmap.

Experimental Environment

The experiment was completed under the ubuntu18.04 operating system. The CPU uses i7-9700K, 8 cores, and 16 threads, with a turbo frequency of 3.6GHz. The GPU uses one NVIDIA 1080Ti. Test in COCO val2017 and COCO test-dev2017.

Experimental Analysis

Table 1 reports the comparison of parameters and calculation accuracy under COCO val2017 and different methods.The unit of AP and AR in the table is the percent sign。

MRNet has an accuracy

of 76.3% when the input size is 256×192, which is better than other methods with the same input size. Compared with the hourglass network[4], the accuracy of the MRNet network is increased by 9.4%. GFLOPs are much lower than the hourglass network, and the amount of parameters is slightly larger. Compared with the CPN network and the CPN+OHKM (online data mining) network[5], the MRNet model has a slightly larger size and a slightly higher complexity, and the accuracy has increased by 7.7% and 6.7%, respectively. Compared with SimpleBaseline[6], MRNet has a significant improvement. Compared with the SimpleBaseline network using ResNet-152 as the backbone network, MRNet's parameters and GFLOPs are greatly reduced, and the accuracy rate is increased by 4.3%. Compared with HRNet using HRNet-W32[7] as the backbone network, MRNet has higher complexity, but its accuracy rate is increased by 1.9%, and the accuracy rate for small-scale targets is also significantly improved. For the use of HRNet-W48[7] As the backbone of HRNet, the accuracy of MRNet has increased by 1.2%, while greatly reducing the number of parameters and calculations. If 384×288 is used as input, compared with HRNetW48, MRNet obtains a 0.4% improvement in accuracy at 71.9% of the calculation cost.

TABLE 1
Comparison of different Methods in COCO val2017

Method	Backbone	Input size	#Params	GFLOPs	AP	AR
8-stage Hourglass[4]	8-stage Hourglass	256×192	25.1M	14.3	66.9	—
CPN[5]	ResNet-50	256×192	27.0M	6.2	68.6	—
CPN+OHKM[6]	ResNet-50	256×192	27.0M	6.2	69.4	—
SimpleBaseline[6]	ResNet-152	256×192	68.6M	15.7	72.0	77.8
HRNet-W32[7]	HRNet-W32	256×192	28.5M	7.10	74.4	79.8
HRNet-W48[7]	HRNer-W48	256×192	63.6M	14.6	75.1	80.4
MRNet(ours)	**MRNet(ours)**	256×192	45.7M	11.3	**76.3**	**80.5**
SimpleBaseline[6]	ResNet-152	384×288	68.6M	35.6	74.3	79.7
HRNet-W32[7]	HRNet-W32	384×288	28.5M	16.0	75.8	81.0
HRNet-W48[7]	HRNet-W48	384×288	63.6M	32.9	76.3	81.2
MRNet(ours)	**MRNet(ours)**	384×288	45.7M	24.5	**76.7**	**81.1**

Table 2 reports the performance comparison between our MRNet and the existing advanced methods in pose estimation. Our method is significantly better than the bottom-up method. Our MRNet network has achieved the highest accuracy rate of 75.9. Under the same input size, our MRNet has an accuracy rate increase of 2.2% compared to Simplebaseline. Compared with HRNet-32 and HRNet-48, our network accuracy rate has been improved. 1% and 0.4%.

TABLE 2
Comparison of different methods in COCO test-dev2017 (#Params and FLOPs are calculated for the pose estimation network, and those for human detection and key-point grouping are not included.)

Method	Backbone	Input size	#Params	GFLOPs	AP	AR
Bottom-up: keypoint detection and grouping						
Openpose[19]	—	—	—	—	61.8	66.5
Associative Embedding[20]	—	—	—	—	65.5	75.4
PersonLab[21]	—	—	—	—	68.7	75.4
MultiPoseNet[22]	—	—	—	—	69.6	73.5
Top-down: human detection and single-person keypoint detection						
CPN[5]	ResNet-Inception	384×288	—	—	72.1	78.5
SimpleBaseline[6]	ResNet-152	384×288	68.6M	35.6	73.7	79.0
HRNet-W32[7]	HRNet-W32	384×288	28.5M	16.0	74.9	80.1
HRNet-W48[7]	HRNet-W48	384×288	63.6M	32.9	75.5	80.5

Mask-RCNN[23]	ResNet-50-FPN	—	—	—	63.1	63.1
G-RMI[24]	ResNet-101	353×257	42.6M	57.0	64.9	69.7
Integral Pose Regression[25]	ResNet-101	256×256	45.0M	11.0	67.8	–
G-RMI+extra data[24]	ResNet-101	353×257	42.6M	57.0	68.5	73.3
CFN[25]	—	—	—	—	72.6	–
RMPE[27]	PyraNet	320×256	28.1M	26.7	72.3	–
MRNet(ours)	**MRNet(ours)**	384×288	45.7M	24.5	**75.9**	**81.3**

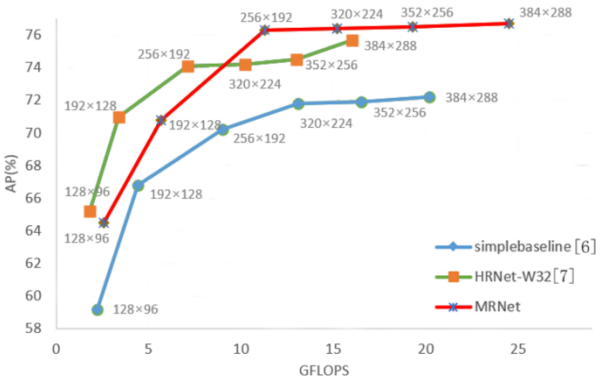
At the same time, to compare the influence of the improved method of the multi-scale feature fusion stage and the feature map output stage on the accuracy of the network model, an ablation experiment was carried out. The experimental results are shown in Table 3. Among them, experiment 1 is the result of HRNet-W32 network test without any improvement. Experiment 2 uses the improved method of MRNet in the multi-scale feature fusion stage to adopt deconvolution and dilated convolution, and no improvement is adopted in other stages. Experiment 3 combines several low-resolution feature maps with high-resolution feature maps only in the feature map output stage. Experiment 4 uses changes in both the multi-scale fusion stage and the feature map output stage, that is, the MRNet network is used. The comparison of experiments 1 and 2 shows that the improvement of the multi-scale fusion stage can improve the accuracy of the network model (increase by 1.6%), and the comparison of experiments 1 and 3 shows that the improvement of the feature map output stage can also improve the accuracy of the network model (increase by 0.2%).), the comparison of experiments 2 and 3 shows that the improvement of the multi-scale fusion stage has a higher effect on the accuracy of the model than the improvement of the feature map output stage, and the comparison of experiments 1 and 4 shows the improvement of the multi-scale fusion stage and the improvement of the feature map output stage. Can jointly improve the accuracy of the network.

TABLE 3
The impact of the improved method on the accuracy of the network model

Experiment	Multi-scale fusion stage improvement	Improved feature map output stage	AP
1			74.4
2	√		76.0
3		√	74.6
4	√	√	76.3

To compare how the input image affects the performance, another six groups of experiments are carried out. The images of different resolutions are input to investigate the sensitivity of the network to images of different scales. The experimental results are compared with the simple baseline network and HRnet network, and the results are shown in Figure 8. It can be seen from the graph analysis that with the increase of the output size, the calculation amount of the network will be greatly improved, and the accuracy will also be slightly improved. Compared with MRnet and the other two methods, when the input image resolution is small, the accuracy of MRnet is close to the other two networks. With the improvement of the input image resolution, the accuracy of MRnet will be higher than the other two networks, which also shows that when the input image resolution is larger, more feature information will be lost in the feature extraction stage. Because of the improvement of MRnet in the multi-scale feature fusion stage, it can better retain the feature information, and ultimately improve the accuracy of the network.

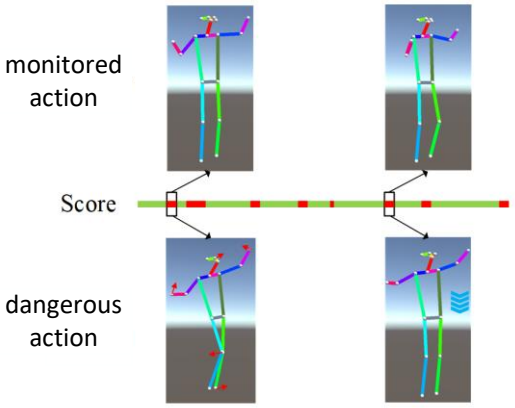
Figure8. The performance of MRnet, HRnet, and simple baseline in different input resolutions, The input resolutions are 128×96、192×128、256×192、320×224、352×256、384×288.



Then, we do experiments on the dangerous action data set. In this dataset, we use each pair of videos as the output to evaluate and match the monitored action with the dangerous action. The experimental results are shown in Figure 9. When there is a big difference between people's behavior and smoking behavior, the guiding

vector will mark some behavior differences. When the action is close to smoking, the score will increase.

Figure 9. According to the action matching diagram, the score of action matching degree in the whole time axis will ultimately affect the result of classification, and the high score will be matched as dangerous action.



CONCLUSION

To complete the task of smoking behavior detection, accurate human posture estimation results are essential. In this paper, a method for human posture estimation, MRnet, is proposed to accurately locate the key points of the human body. In the multi-scale feature fusion stage, a new convolution operation is introduced to retain more feature information and reduce the loss of speech information in the whole feature extraction process. In the output stage of the feature map, the semantic information of low-resolution features is combined to get a high-quality heat map with richer semantic information, to achieve more accurate human pose estimation. In the experiment, OKS is used to evaluate the accuracy. The experimental results show that MRnet has higher accuracy than the mainstream human pose estimation methods. At the same time, it plays an important role in smoke detection and artificial intelligence action recognition.

Future work should focus on the indoor environment with the dense crowd, occlusion, and less light. To detect the smoking behavior in the production and work area, to achieve the purpose of tobacco control and safe production. At the same time, we should also pay attention to the improvement of the lightweight model. By modifying the model, pruning the model, and compressing the model, we can lighten the network, reduce the calculation cost on the premise of ensuring the performance of the model and detect and alarm the smoking behavior in a large number of monitoring devices.

DECLARATIONOFCONFLICTINGINTERESTS

Theauthor(s)declarednopotentialconflictsofinterestwithrespecttotheresearch,author-ship,and/orpublicationofthisarticle.

FUNDING

This work was supported, the Beijing Municipal Commission of Education Project (No.KM202111417001, KM201911417001), the National Natural Science Foundation of China (Grant No. 61871039, 61906017, 61802019), the Collaborative Innovation Center for Visual Intelligence (Grant No. CYXC2011), the Academic Research Projects of Beijing Union University(No. ZB10202003, ZK40202101, ZK120202104, XP202015), the postgraduate research and innovation funding project of Beijing Union University (YZ2020K001).

REFERENCES

1. Li M, Wei F, Li Y, et al. Three-Dimensional Pose Estimation of Infants Lying Supine Using Data from a Kinect Sensor With Low Training Cost. in IEEE Sensors Journal 2021; 21: 6904-6913.
2. Jiao Y, Yao H, and Xu C. PEN: Pose-Embedding Network for Pedestrian Detection. in IEEE Transactions on Circuits and Systems for Video Technology 2021;31: 1150-1162.
3. Zhang Y, Guan S, Cheng Xu, et al. Based on Spatio-temporal Graph Convolution Networks with Residual Connection for Intelligence Behavior Recognition. International Journal of Electrical Engineering & Education 2021: 1-15.
4. Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation. European Conference on Computer Vision 2016: 483-499.
5. Chen Y, Wang Z, Peng Y. Cascaded pyramid network for multi-person pose estimation. IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.
6. Xiao B, Wu H, Wei Y. Simple Baselines for Human Pose Estimation and Tracking. European conference on computer vision 2018: 466-481.
7. Sun K, Xiao B, Liu D. Deep High-Resolution Representation Learning for Human Pose Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019: 5693-5703.
8. Yu Y, Li H, Cao J, et al. Three-Dimensional Working Pose Estimation in Industrial Scenarios with Monocular Camera. in IEEE Internet of Things Journal 2021;8: 1740-1748.
9. Valle R, Buenaposada J M, A Valdés, et al. Face Alignment using a 3D Deeply-initialized Ensemble of Regression Trees. Computer Vision and Image Understanding 2019;189: 102846.
10. Zhou Y, Hu X, Zhang B. Interlinked Convolutional Neural Networks for Face Parsing International Symposium on Neural Networks. Springer, Cham 2015: 222-231.
11. Liu H Z, Deng Z F. Learning Spatiotemporal Features with 3D DenseNet and Attention for Gesture

Recognition. International Journal of Electrical Engineering Education 2019: 1-18.
12. Hu Q, Tang X, Tang W. A Smart Chair Sitting Posture Recognition System Using Flex Sensors and FPGA Implemented Artificial Neural Network. in IEEE Sensors Journal 2020;20: 8007-8016.
13. Feng L, Li Z, Liu C, et al. SitR: Sitting Posture Recognition Using RF Signals, in IEEE Internet of Things Journal 2020;12: 11492-11504.
14. Ren W, O Ma, H. Ji, et al. Human Posture Recognition Using a Hybrid of Fuzzy Logic and Machine Learning Approaches. in IEEE 2020;8: 135628-135639.
15. Cheng B, Xiao B, Wang J, et al. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020: 5385-5394.
16. F. Li, H. Zhou, Z. Wang, et al. ADDCNN: An Attention-Based Deep Dilated Convolutional Neural Network for Seismic Facies Analysis with Interpretable Spatial–Spectral Maps. in IEEE Transactions on Geoscience and Remote Sensing 2021;59: 1733-1744.
17. Lin T Y, Maire M, Belongie S. Microsoft coco: Common objects in context. European conference on computer vision. Springer, Cham 2014: 740-755.
18. V. Hoang and K. Jo. 3-D Human Pose Estimation Using Cascade of Multiple Neural Networks. in IEEE Transactions on Industrial Informatics 2019; 15: 2064-2072.
19. Cao Z, Gines Hidalgo, Tomas Simon, et al. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell 2021;43: 172-186.
20. A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In NIPS 2017: 2274–2284.
21. G. Papandreou, T. Zhu, L.C, et al. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In ECCV September 2018: 269-286.
22. M. Kocabas, S. Karagoz, E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In ECCV Lecture Notes in Computer Science 2018;11215: 437–453.
23. Wu Q F, Feng D Q, Cao C Q, et al. Improved Mask R-CNN for Aircraft Detection in Remote Sensing Images. Sensors 2021;21(8) 2618.
24. G. Papandreou, T. Zhu, N. Kanazawa, et al. Towards accurate multi-person pose estimation in the wild. In Conference on Computer Vision and Pattern Recognition 2017: 3711–3719.
25. Sun X, Xiao B, Wei F, et al. Integral human pose regression. In ECCV 2018: 536–553.
26. Huang S, Gong M, Tao D. A coarse-fine network for keypoint localization. In ICCV, IEEE Computer Society 2017: 3047–3056.
27. Fang H, Xie S, Tai Y, et al. RMPE: regional multi- person pose estimation. In ICCV 2017: 2353 − 2362.