Classroom Student Posture Recognition Based on an Improved High-Resolution Network

Yiwen Zhang
University of South China
Tao Zhu (tzhu@usc.edu.cn)
University of South China
Huansheng Ning
University of Science and Technology Beijing
Zhenyu Liu
University of South China

Research

Keywords: Pose Estimation, Support Vector Machine, High-Resolution Networks, Squeeze-and-Excitation Networks, Object Detection

Posted Date: September 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-72287/v1

Abstract

Due to the large number of students in a university classroom and crowded seating, most features are obscured, making it difficult to balance accuracy in identifying students' postures with computational speed. Aiming at the above problems, a classroom student postures recognizes method is proposed. First of all, because we need to recognize multi-person poses in the classroom, we use you only look once (YOLOv3) as the object detection algorithm, and retrain the YOLOv3 to detect human object that groveling to the table. Then to improve the accuracy of pose estimation network, we use the Squeeze-and-Excitation Networks (SENet) structure embedded in the residual structure of High-Resolution Networks (HRNet). Finally, using the improved HRNet algorithm output of human body key points, we designed a pose classification algorithm based on Support vector machine (SVM), which is used to classify human poses in the classroom. Experiments show that the improved HRNet multi-person poses estimation algorithm result can reach the best AP performance of 73.76 points in the common objects in context (COCO) validation dataset. We tested our designed posture classification algorithm using our own dataset collected in the classroom and were able to achieve a high recognition rate (90.1%) and robustness. and it could effectively solve the difficulty of student postures recognition.

Introduction

In recent years, benefited from the growth of surveillance systems for both public and personal usage, pose estimation and detection methods have been greatly developed to meet the emerging needs of various industries. There are many problems behaviors that university students have in class, such as sleeping, playing mobile phones and chatting. These inappropriate behaviors will affect students' classroom learning. classroom learning efficiency is one of the important factors affecting their academic performance. Therefore, in the university classroom, students pose estimation and detection are the application of computer vision technology in the field of university education, it has very important research significance and application value.

There are two candidates for this task. One is based on object detection algorithms. The paper [1][2] used the improved Faster R-CNN [3] model as the object recognition to detect student postures in classroom. And the paper [4] only detects students sleeping pose base on improved R-FCN [5]. These methods can detect poses in small, low-quality pictures of classroom, where students are concentrated and the collected pictures have low resolution. However, using object detection methods to estimate poses of each individual on this occasion are hindered by two main obstacles. First, each method is low scalability. If we need to detect a new pose, it has to re-train the entire network. Secondly, those methods only identified pose that significantly differed from others, such as sitting, standing. For other less distinct pose, such as: reading, chatting and raising your hand, are generally cannot recognized.

The other method is based on the pose estimation network. The paper [6] uses OpenPose [7] to estimate the location of human body keypoints, and then use a classifier to classify the collected keypoints. The advantage of this method is easy to implement and it has fast calculation speed, but the disadvantage is

low accuracy of the calculation results. The paper [8] uses pose estimation maps(heatmaps), the byproduct of pose estimation, to recognize human action. This method is very effective when applied to behaviors with large movements, but it is difficult to learn behaviors with small movements.

Due to the large number of students in classrooms, and human bodies covered by objects such as tables or other human bodies, there are four main obstacles in recognizing students' posture in classroom.

1. The estimation of the human body keypoints has a high error rate and low accuracy rate.
2. Part of human joints is invisible to camera due to heavy occlusions, relying on only a few unreliable features to estimate human body keypoints. So, it is hard to recognize the grovel posture.
3. Most of the Top-Down pose estimation methods calculation speed is low, so the results cannot be displayed in real time.
4. If we only use object detection method to detect pose in classroom, there will be some problems such as it can only detect a single gesture or it will have poor scalability.

Faced with these difficulties, we propose a pose recognition method in classroom which combines the pose estimation algorithm and the object detection algorithm. Our contributions are three-fold as follows: (1) We capitalize on YOLOv3 (You Only Look Once) model [9] as the object detection algorithm to detect human object and student groveling on the table object. (2) We propose improved HRNet model as the pose estimation algorithm to solve the problem of high error rate of estimate human body keypoints. We term our improved HRNet as SE-HRNet. SE-HRNet model is achieved by embedding the SENet [10] structure into the HRNet model [11]. (3) Finally, we design a multi-posture classification network base on Support Vector Machine (SVM) [12]. Experimental results show, the AP that uses YOLOv3 to detect the Groveling pose is 91.6 points, the AP that uses the SE-HRNet model to detect keypoints of the human body is 73.7 points, the accuracy of the pose classification is 88.6%, and the computation speed of our classroom student postures recognize method is 7 images per second.

The rest of this paper is organized as follows. In Section 2, we will introduce the related work. And then proposes the classroom student postures recognize method in Section 3. Finally, in Section 4 we introduce our dataset and then discuss our experimental results in detail. Conclusion is summarized in Section 5.

2. Related work

2.1 Pose estimation methods

At this stage, multi-person human poses estimation methods are divided into two categories:

Top-Down approaches: First, perform object detection of the human body on the image and crop. Then use the single-person pose estimation for each cropped human body. So, for each detection, a single-person pose estimator is run, and the more people there are, the greater the computational cost. but the

accuracy of the Top-Down method is usually relatively higher. The common models include CPN [13], Hourglass [14], CPM [15], Alpha Pose [16] and so on.

Bottom-Up approaches: First it detects all the keypoints of the human body in the picture, and then matches these points to different individuals, so this method is faster in calculation, but the accuracy is slightly lower than Top-Down method. The common model such as: OpenPose [7].

High-resolution network (HRNet) [11] is a state-of-the-art human body pose estimation method. And a Top-Down method. High-resolution network pose estimation is able to maintain high resolution representations through the whole process. It starts from a high-resolution subnetwork as the first stage, gradually adding high-to-low resolution subnetworks one by one to form more stages, and connecting the multi-resolution subnetworks in parallel [11]. It performs multiple multi-scale fusions by repeatedly exchanging information across parallel multi-resolution subnetworks through the whole process. It estimates the keypoints of the human body through the high-resolution representations of the network output. The architecture of the HRNet is illustrated in **Figure. 1**.

It has two benefits in comparison to the common pose estimation networks [13-16]. First, this approach connects high to low resolution subnetworks in parallel, rather than serial, as most existing solutions do. Therefore, it is able to maintain high resolution rather than restoring resolution through a low to high process. and accordingly, the predicted heatmap is spatially more precise. Second, most existing fusion schemes combine low-level and high-level representations [11]. On the contrary, this method uses the low-resolution representation of the same depth and similar level to perform multiple multi-scale fusion to improve the high-resolution representation, and vice versa, making the high-resolution representation also rich in pose estimation. As a result, this method predicted heatmaps are more accurate.

HRNet can maintain high-resolution features through the whole process without the need of recovering the high resolution. It is also fuse parallel multi-resolution representations repeatedly, enhance reliability of high-resolution representations. It yields accurate and spatially precise point heatmaps. But because HRNet is a Top-Down method, its image processing speed will be slower than Bottom-Up method. In addition, in order to realize HRNet multi-person pose estimation, the object detection algorithm needs to process the image first. Therefore, the detection speed of object detection algorithm has great influence on the speed of pose estimation.

2.2 Squeeze-and-Excitation Networks

Squeeze-and-Excitation Networks (SENet) [10]. SENet introduced a new architectural unit, which term the Squeeze-and-Excitation (SE) blocks, with the goal of improving the quality of representations produced by a network by explicitly modelling the interdependencies between the channels of its conventional features [10]. In this structure, Squeeze and Excitation are two very critical operations. A new "feature recalibration" strategy is adopted, through this mechanism networks can learn to use global information to selectively emphasize informative features and suppress less useful features.

The structure of the SE building block is shown in **Figure. 2**. Among them: SE represents the SENet blocks. The first passed through a squeeze operation, which first performs global average pooling on the input feature map to obtain a feature map of size C×1×1 (C is the number of feature map channels), allowing information from the global receptive field of the network to be used by all its layers. The aggregation is followed by an excitation operation, through the parameter is used to generate weight for each feature channel, where the parameter is learned the correlation between the feature channels. After two fully connected layers (first dimensionality reduction and then dimensionality increase), it uses the sigmoid activation function to obtain a weight of C×1×1. Finally, there is a reweight operation. We regard the output weight as the importance of each feature channel after feature selection, and then weight the previous features one by one through multiplication to complete the feature recalibration. The output of the SE blocks which can be fed directly into subsequent layers of the network. Both BasicBlock and Bottleneck are the classic residual modules used in the ResNet network. SE-BasicBlock means to embed the SE structure into the regular BasicBlock unit, and SE-Bottleneck means to embed the SE structure into the regular Bottleneck unit.

The structure of the SE block is simple and can be directly embedded into the existing state-of-the-art network architectures, which has a significant improvement in the results of network, also computationally lightweight and impose only a slight increase in model complexity and computational burden.

2.3 Object detection method

The existing object detection algorithms are mainly divided into two types, the two-stage method (region proposal method) and the one-stage method (regression method). The common two-stage object detection algorithms include RCNN [18], Fast RCNN [19] and Faster-RCNN [20]. Among them, Faster R-CNN tends to be a slower but more accurate model [21]. Faster R-CNN consists of two stages. In the first stage, called the region proposal network (RPN). Images are processed by RPN to predict class-agnostic box proposals. In the second stage, these proposals boxes is used to crop features from the same intermediate feature maps which are then entered into the feature extractor in order to predict a class for each proposal box and to optimize for each proposal box.

In one-stage object detection algorithm, e.g., SSD [22] and YOLO [23], conduct object classification and bounding-box regression concurrently without a region proposal stage [24]. YOLO converts object detection into regression work. Based on a single end-to-end network, complete the calculation from the original image to the output of the object position and category. These one-stage methods usually have a high detection speed and high efficiency but low accuracy. YOLOv3 [9] can detect multiple objects with a single inference, so its detection speed is therefore extremely fast; in addition, by applying a multi-stage detection method, it can complement the low accuracy of YOLO and YOLOv2 [25]. Although YOLOv3 is not as good as Faster-RCNN in terms of detection accuracy for very small targets. But from the perspective of detection speed, YOLOv3 is significantly better than Faster-RCNN [20]. So YOLOv3 is suitable for many engineering applications.

Considering the detection speed and detection accuracy of the algorithm, this paper uses the YOLOv3 as the object detection algorithm to detect human body and groveling poses in the classroom, and provides the foundation for our real-time classroom human pose recognize method based on SE-HRNet.

Methodology

3.1 Overview of the framework

The overview of the classroom student postures recognize method proposed in this paper is shown in Figure. 3. First, we use pre-trained YOLOv3[9] to detect the images we collect from classrooms. The output results fall into two categories, one is the human body object provided for SE-HRNet pose estimation, and the other is the Groveling pose object. Then the results of the Groveling pose object are directly output, and the results of the human body object are cropped from the image. The cropped images are input into SE-HRNet for pose estimation. SE-HRNet detects the locations of 17 keypoints of the human body. The next step is to preprocess the data of output keypoints. We design an SVM classifier to classify the preprocessed keypoints of the human body. Then output the classification results. Finally, the proposed method applied to online detection of real surveillance images of the classroom.

3.2 YOLOv3 application

Because of the limitations of classroom usage scenarios, the results of student postures recognize need to be displayed in real time and it needs to be as accurate as possible. So, we need to deal with the slow estimation speed of HRNet which is a top-down pose estimation method. The original object detection network used by HRNet is Faster R-CNN [20]. Through our discussion in the 2.3 section of this paper. We propose to replace Faster R-CNN with YOLOV3 [9] as the object detection network of our classroom student postures recognize method.

Among the three poses we proposed to recognize, the groveling pose is the hardest to recognize and classifier through the pose estimation network. Because most of the features of the groveling pose have been lost, if we try to use the pose estimation network to estimate the groveling pose, the human body keypoints of the groveling pose will be seriously lost. Which makes the pose classifier impossible to classify the groveling pose.

To solve the above problem, the most effective method is to use the object detection network to detect the groveling pose. We also need to use the object detection network to detect human objects in the classroom for the pose estimation network. So, we are using the datasets we collected to retrain the YOLOv3 to detect the groveling pose, and improve the accuracy of human object detection in classroom.

3.3 Design The improved HRNet

When the human body overlaps with each other, many human body features will occlude in images. Especially in crowded and complex places, such as classrooms. Conventional pose estimation networks

output feature maps have high confidence in the keypoints of the overlap's parts. The network mistakenly "believes" that the overlaps or missing keypoints are also part of the human body. This unbalanced confidence distribution causes a large number of misidentifications [17]. For further analysis, in order to enable the network to learn more global features, the method of enhancing the receptive field can be adopted to balance the confidence of the heat map in different positions. Therefore, we propose to embed the SENet structure into HRNet to increase the global information of HRNet.

The Squeeze operation in the SENet structure converts a feature map into a number, which has a global receptive field, and two full connection layers serve to reduce the number of parameters. HRNET for feature extraction is the key to accurately estimate the keypoints, which the residual layer fused multiple layers feature. Therefore, we propose embeds the SENet structure into BasicBlocks and Bottleneck of the HRNet to obtain the SE-BasicBlock and SE-Bottleneck substructure see **Figure. 2**, thereby expanding the receptive range of the feature map to global information.

The structure of SE-HRNet is shown in **Figure. 4**. SE-HRNet consists of four stages with four parallel subnetworks, the resolution is gradually reduced to half and the width (number of channels) is correspondingly increased twice. This paper embeds SENet structure into the first stage (Stagel1), which contains 4 SE-Bottleneck units, is composed of a SE-Bottleneck with the width 64, and is followed by one 3×3 convolution feature map to reduce the width to C (the number of channels C), the second, third, and fourth stages contain 1, 4, and 3 exchange blocks respectively. One exchange block contains 4 SE-BasicBlock embedded in SENet structure at each resolution, where each contains two 3×3 convolutions, and exchange units across resolutions. In summary, there are a total of 8 exchange units, i.e., a total of 8 multi-scale fusions are conducted.

The SE structures introduces primitive information into deep layers, inhibits information degradation, and then expands the receptive field by pooling. It integrates shallow information and deep information from multiple dimensions, so that the combined output contains multiple levels of information, which enhances the feature map expression ability.

3.4 Design The classification method

Data Preprocessing.

In order to reduce the amount of calculation, speed up the convergence, and improve the accuracy, the human body keypoints data output from SE-HRNet needs to be preprocessed. First, because the coordinate origin of the keypoints data output by SE-HRNet is in the upper left corner of the image, and each image contains multiple human bodies, so first it is necessary to shift the coordinates origin to the nose position of the 17 points in each human body. Then we normalize the data, scale the coordinate data to between 0 and 1 according to the image resolution.

Design SVM classifier.

The classifier structure is a simple four layers fully connected network. Each layer has 125 neurons and uses Rectified Linear Unit (ReLU) as the activation function. We use the Adam optimizer. The base learning rate is set as 1e-3. The training process is terminated within 150 epochs. We only use classifier to classify two types of actions in classroom: reading and looking. and the loss function used is the hinge loss see Eq. (4) form of Support Vector Machine (SVM) [12]. The simplest way to extend SVMs for multiclass problems is using the so-called one-vs-rest approach [26].

$$\min_{w} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max(1 - w^T x_n t_n, 0)^2 \qquad (1)$$

Because classroom student posture recognize method combines object detection, pose estimation and keypoints classification, and the keypoints classification network. So, if we want to recognize a new pose in the classroom, we only need to retrain the keypoints classification network, instead of retraining all the networks of the entire method. Therefore, improving the scalability and practicability of the method.

Experimental

4.1 Dataset and setup

Dataset.

In this experiment, the dataset for training and validating the improved HRNet is COCO2017 [27], which includes 149,808 pictures, and over 250,000 person instances labeled with 17 key points. We evaluate our improved HRNet on val2017 set, containing 6384 pictures.

In order to reflect the situation in the real classroom environment as much as possible, we collected a dataset in this paper. It contains a large number of images with different degrees of occlusion and light changes (see **Figure. 6**). Our dataset was taken by 10 Dahua network dome surveillance cameras installed in 6 classrooms of the University. The picture resolution is 2592×1520. The data collected in this paper contains 1,951 training samples and 943 test samples. Each picture has an average of 5 people. We labeled a total of 14,470. There are three types of poses annotated in this dataset including Reading, Groveling and Looking, as shown in **Figure. 5**

Experimental environment.

The software environment of experiment is Ubuntu 18.04 based on PyTorch 1.4 with CUDA 10.1, the hardware environment is Intel Core i7 7820X CPU, 64GB RAM, and the graphics card is Nvidia TITAN X (Pascal) 12G.

Evaluation Metrics.

The standard evaluation metric of the pose estimation experiment is based on Object Keypoint Similarity (OKS) in Eq. (5). Here is the Euclidean distance between the detected keypoints and the corresponding ground truth, is the visibility flag of the ground truth, is the object scale, and is a per-keypoint constant that controls falloff. We report standard average precision and recall scores[11]: Average Precision (AP), stands for the mean of AP scores at 10 positions, OKS = 0.50, 0.55...0.9, 0.95, and the same goes for Average Recall (AR).

The evaluation measurements of the object detection algorithm are Average Precision (AP), it was proposed in [28]. a common judgment for the correctness is the intersection-over-union (IOU) between detection result and ground truth. If the IOU is greater than a threshold percentage of the ground truth size, the result is considered correct [1]. In order to get a higher recall, we set the IOU threshold to 0.5.

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

Results And Discussion

5.1 Train YOLOv3 on our dataset

YOLOv3 uses Darknet-53 backbone, and pretrained the backbone on the COCO dataset. We set network input resolution to 416×416. We use multi scale training. The dataset we use to retrain and test YOLOv3 is the data of real classroom images that we collect. We retrain YOLOv3 to detect the groveling pose and student body. The training process is finished within 150 epochs. After retrain YOLOv3, the best result on the test set for Groveling pose are AP=91.6 points.

Table 1. Comparisons of different methods on the COCO validation set

Method	input size	#Params	GFLOPs	AP	AR
OpenPose [7]	368x368	—	—	61.8	66.5
Baseline ResNet-50[29]	256×192	34.0 M	8.90	70.4	76.3
HRNet-W32(paper) [11]	256×192	28.5 M	7.10	73.4	78.9
HRNet-W32(our implement)	256×192	28.54 M	7.20	73.1	78.7
SE-HRNet-W32(our)	256×192	28.75 M	7.21	73.8	79.2

5.2 Compare different pose estimation methods

In order to verify the effectiveness of the improved HRNet, several pose estimation frameworks have been compared, including OpenPose [7], original HRNet [11], and ResNet [29].

Both original HRNet [11] and SE-HRNet trained on COCO2017 [27] from scratch with the input size 256×192. The learning rate and Dropout rate are unchanged according to the settings in the original paper [11]. The training is set for a total of 210 epochs.

The **Table 1** shows the results of our improved HRNet compare to other multi-person pose estimation methods on the COCO verification set. our improved HRNet by embedding the SENet structure, achieves a 73.7 AP score, outperforming other methods with the same input size (256×192) except OpenPose[7], OpenPose input size is 368x368, and it is a bottom-up approaches. Our approach is much better than bottom-up approach, compared to the OpenPose our improved network improves AP by 11.9 points. And compared to the SimpleBaseline-ResNet-50 [29], our obtains significant improvements: gain 3.3 points with a smaller model size and smaller GFLOPs.

From the results, our HRNet [11] training results showed a slight decrease (0.3 point) compared to the training results provided in the paper [11], also the GFLOPs and the number of parameters is slightly larger. The SE-HRNet that the HRNet network embedded with the SENet structure compared to original HRNet that we trained has 0.7 points improvements, model size (#Params) and GLOPs did not increase significantly, both increase within 1%.

5.3 Compare different methods

To verify the effectiveness of the proposed method, we try to combine different pose estimation and object detection algorithms with pose classification algorithms. The results are shown in **Table 2**. First, we tried to use OpenPose +SVM as the classroom student postures recognize method. However, the OpenPose pose estimation is not accurate enough, so a large number of classify errors occur, and the groveling pose cannot be recognized. Because the human body features of the groveling pose are basically occlusions, OpenPose cannot output any useful human body keypoints. Second, the method using Faster RCNN +HRNet +SVM has some improvements over the method using OpenPose, and it can recognize the groveling pose because we use the Faster RCNN to detect the groveling pose. Also, the accuracy of Faster RCNN detects the groveling pose is quite high. Due to the Faster RCNN is a two-stage object detection algorithm. Finally, the YOLOv3 +SE-HRNet +SVM that we proposed in this paper has significantly improvements (8.3%) compared to other methods, can reach 90.1%. Although YOLOv3 is an one-stage object detection algorithm, the accuracy of detecting the groveling pose is similar to Faster RCNN.

Table 2. Comparisons of different methods on our dataset.

Method	Reading	Looking	Groveling	Accuracy
OpenPose +SVM	67.3%	61.4%	—	64.2%
Faster RCNN +HRNet-W32 +SVM	83.4%	84.5%	92.4%	81.8%
YOLOV3 +SE-HRNet-W32 +SVM (our)	88.6%	89.2%	91.6%	90.1%

5.4 Computation cost

To evaluate the computational cost, we tested different approaches on a PC with the same running configuration, and the configuration is in Experimental environment. The result is shown in **Table 3**. None of the methods add the last step pose classification.

It can be seen that the average running time of each image in our method can reach 0.142s. Our method is a bit slower than OpenPose, because OpenPose is a bottom-up method. But our method is very close to the HRNet + YOLOV3 method and significant faster than HRNet-W32+ Faster RCNN by 404%. This also shows that our improved HRNet does not add much computational cost with the addition of the SENet structure. This means that our approach works very well in practical applications in classroom environments.

Table 3. Compare the processing time of each image of these two methods on our dataset

Methods	Time (s)	Frames Per Second (FPS)
OpenPose	0.11	10
HRNet-W32+ Faster RCNN	0.321	3
HRNet + YOLOV3	0.136	7
SE-HRNet + YOLOV3(our)	0.142	7

5.5 Discussion

The results of SE-HRNet tested on our dataset are compared with the OpenPose, as shown in **Figure. 6**. OpenPose mistakenly identified the patterns of the wall as human bodies and the human pose estimation accuracy is not good. SE-HRNet compares to OpenPose obtain significant improvements: reduced the human body object detected error rate and increased the accuracy of estimate keypoints.

Figure. 7 shows the representative recognized results of our proposed method from sparse to dense situations. In which each pose is labeled right next to body keypoints, and groveling posture is labeled the bounding-box in the images. It can be seen that our method can deal with the sparsely distributed and concentrated distribution of students. Our method can clearly locate these students and recognize their pose in difficult situations where students are crowded and occlude each other. In another difficult situation, where some grovel posture can easily be confused with the looking pose, our approach still predicts the correct label. In other cases, such as occlusion or background noise, our method also shows it is more robustness than other existing methods.

There may be some possible limitations in this study. Due to our limited manpower, we collect a small amount of data tagged in the dataset. The amount of data used to test our proposed methodology is not all large enough to be a good representation of all classroom environments.

Conclusions

In this paper, we propose a classroom student postures recognize method. The proposed method is combined the pose estimation, object detection and postures classification to complete the classroom postures recognition. This paper compares the speed and accuracy of different methods, then chose YOLOv3 as our object detection network to detect the grovel postures. Then because the high error rate of other common pose estimation methods, we propose to embed the SENet structures into HRNet. Experiment shows, we tested our improved HRNet on the COCO dataset. The AP reaches 73.8 points, compare to the original HRNet gains 0.7 points. And the GFLOPs and #Params have slightly increased. Finally, we design a postures classification algorithm based on SVM. The accuracy of our proposed method can reach 90.1% outperforms other tradition methods. And this method combined three different modules in order to recognize student postures, so it can have strong robustness and scalability. In future work, we will redesign the posture classification algorithm to recognize more student postures. And labeling more data from different students in different environments and at different times will further improve the adaptability of our approach.

Declarations

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that there is no conflict of interest.

Funding

This work is supported by the National Natural Science Foundation of China (No. 61872038), Natural Science Foundation of Hunan Province (No. 2019JJ50499)

Authors' contributions

Yiwen Zhang performed the experiments and was a major contributor in designing the new method and writing the manuscript, Tao Zhu contributed to the proposed methods and the experiments. Huansheng Ning and Zhenyu Liu funded for and conceived the idea of the work. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Abbreviations

HRNet: High-Resolution Network

SENet: Squeeze-and-Excitation Networks

YOLO: You Only Look Once

SVM: Support vector machine

COCO: common objects in context

References

1. L. Tang, C. Gao, X. Chen, Pose detection in complex classroom environment based on improved Faster R-CNN. IET Image Processing. **13**(3), 451–457 (2019)
2. T. Bin, Y. Shu-Han, Research on the Algorithm of Students' Classroom Behavior Detection Based on Faster R-CNN. Modern Computer. (2018)
3. S. Ren, K. He, R. Girshick, R.-C.N.N. Faster, Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence. **39**(6), 1137–1149 (2017)
4. W. Li, F. Jiang, R. Shen, in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019 IEEE Conference. Sleep Gesture Detection in Classroom Monitor System (Brighton, 2019), pp. 7640–7644
5. J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks. Advances in Neural Information Processing Systems, 379–387 (2016)
6. J. Zaletelj, in Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis. Estimation of students' attention in the classroom from kinect features (Ljubljana, 2017), pp. 220–224
7. T. Zhe Cao, S.-E. Simon, Y. Wei, Sheikh, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields (Hawaii, 2017), pp. 7291–7299
8. J. Mengyuan Liu, Yuan, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2018 IEEE Conference. Recognizing Human Actions as the Evolution of Pose Estimation Maps (Salt Lake, 2018), pp. 1159–1168
9. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement. (arXiv, 2018)https://arxiv.org/abs/1804.02767. Accessed 8 April 2018
10. Hu Jie, L. Shen, Gang Sun, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference. Squeeze-and-Excitation Networks (Salt Lake, 2018),

pp. 7132–7141

11. S. Ke, B.X.D. Liu, J. Wang, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019 IEEE Conference. Deep High-Resolution Representation Learning for Human Pose Estimation (Long Beach, 2019), pp. 5693–5703
12. B.E. Boser, A training algorithm for optimal margin classifiers. Paper presented at ACM Fifth Workshop on Computational Lerning Theory, Pittsburgh, 1992
13. Y. Chen, Z. Wang, Y. Peng, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference. Cascaded Pyramid Network for Multi-person Pose Estimation (Salt Lake, 2018), pp. 7103–7112
14. A. Newell, K. Yang, J. Deng, Stacked Hourglass Networks for Human Pose Estimation. by Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9912 (Springer, Cham,2016), pp. 483–499
15. S.E. Wei, V. Ramakrishna, T. Kanade et al., in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference. Convolutional Pose Machines (Las Vegas, 2016), pp. 4724–4732
16. IEEE International Conference on Computer Vision (ICCV)
H. Fang, S. Xie, Y. Tai, C. Lu, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017 IEEE Conference. RMPE: Regional Multi-person Pose Estimation (Venice, Italy, 2017), pp. 2334–2343
17. L. Xueping Liu, L. Yuqian, Li, Improved YOLOV3 Object Recognition Algorithm with Embedded SENet Structure. Computer Engineering. **45**(11), 243–248 (2019)
18. J.R.R. Uijlings, T. Sande K E A V D, Gevers, Selective Search for Object Recognition. Int. J. Comput. Vision **104**(2), 154–171 (2013)
19. IEEE International Conference on Computer Vision (ICCV)
R. Girshick, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015 IEEE Conference. R.-C.N.N. Fast (Santiago, Chile, 2015), pp. 1440–1448
20. S. Ren, K. He, R. Girshick et al., R.-C.N.N. Faster, Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **39**(6), 1137–1149 (2017)
21. J. Huang, V. Rathod, Chen Sun, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors (Hawaii, 2017), pp. 7310–7311
22. W. Liu et al.: SSD: Single Shot MultiBox Detector. by Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905 (Springer, Cham,2016)
23. J. Redmon, S. Divvala, R. Girshick, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference. You Only Look Once: Unified, Real-Time Object Detection (Las Vegas, 2016), pp. 779–788

24. /CVF International Conference on Computer Vision (ICCV)
D. Jiwoong Choi, H. Chun, Kim, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019 IEEE Conference. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving. (Seoul, Korea, 2019), pp. 502–511
25. J. Redmon, A. Farhadi, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference. Yolo9000: better, faster, stronger (Hawaii, 2017), pp. 7263–7271
26. Deep Learning using Linear Support Vector Machines
Y. Tang, Deep Learning using Linear Support Vector Machines. (arXiv, 2013), https://arxiv.org/abs/1306.0239. Accessed 2 June 2013
27. T. Lin, M. Maire, S.J. Belongie, Microsoft COCO: common objects in context (2014), https://cocodataset.org. Accessed 2014
28. R. Padilla, S.L. Netto and E. A. B. da Silva, in the 27th International Conference on Systems, Signals and Image Processing (IWSSIP). A Survey on Performance Metrics for Object-Detection Algorithms (Online, 2020), pp. 237–242
29. B. XIAO, H. WU, WEI Y., in 15th European Conference on Computer Vision (ECCV). Simple Baselines for Human Pose Estimation and Tracking (Munich, Germany, 2018), pp. 472–487

Figures

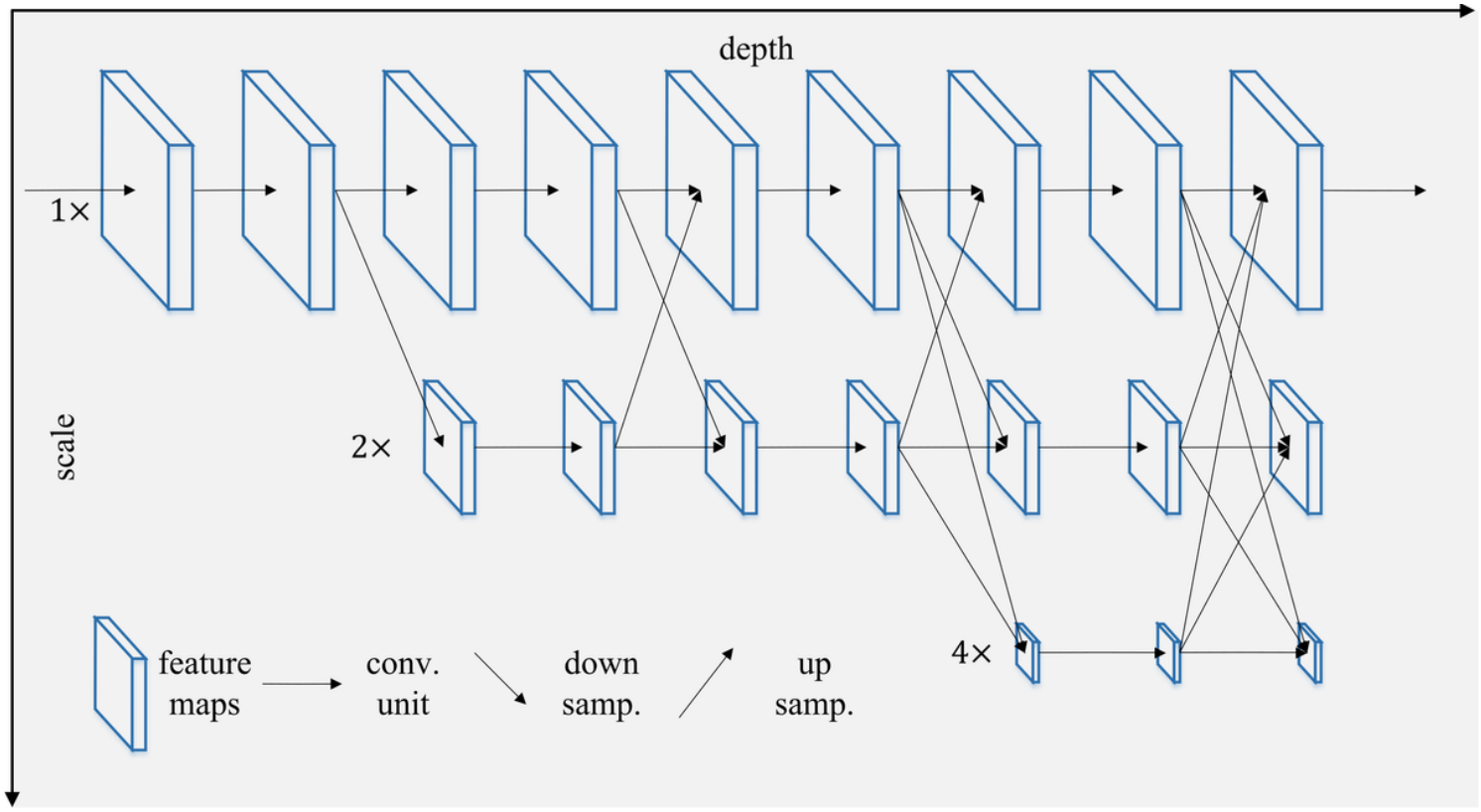

Figure 1

The architecture of the HRNet. The architecture of the HRNet. It consists of parallel high-resolution and low-resolution subnet-works with repeated information exchange between multi-resolution subnetworks (multi-scale fusion).

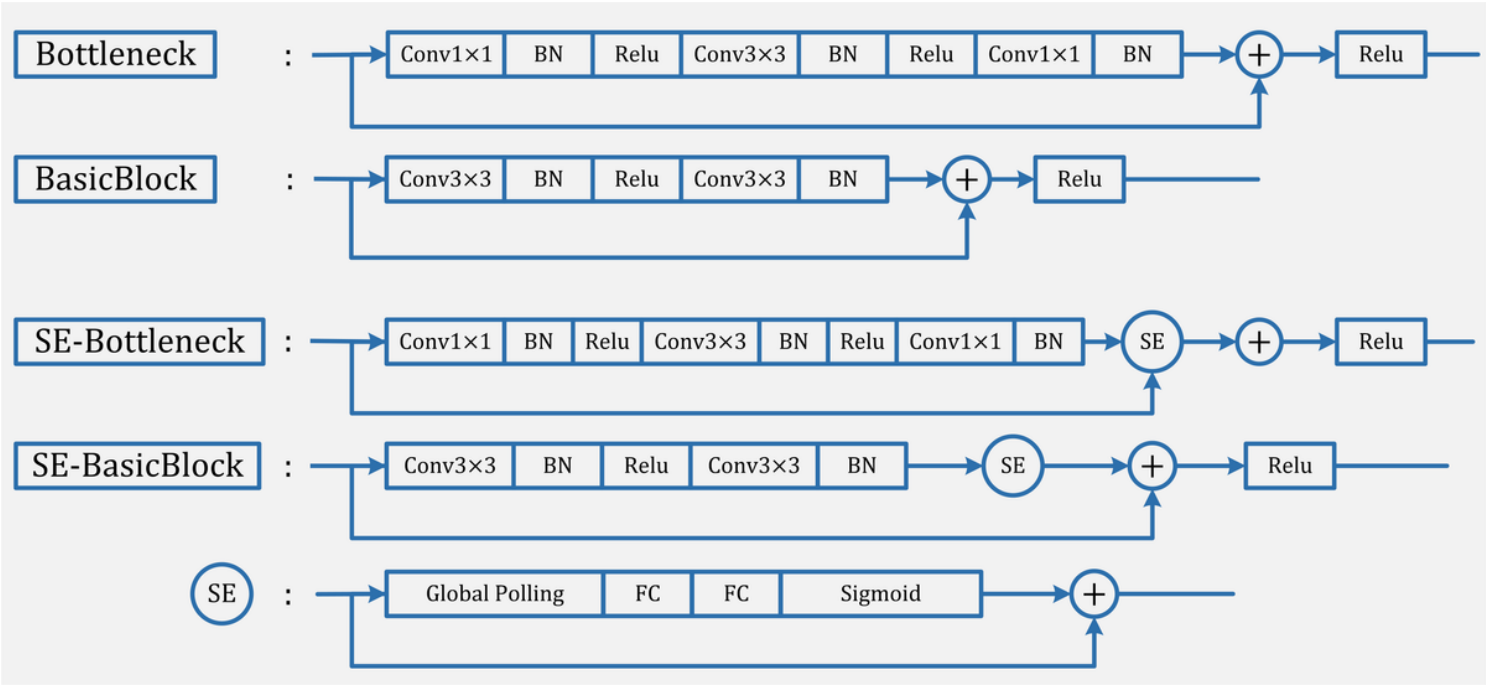

Figure 2

The architecture of SENet. BasicBlock and Bottleneck are original Residual modules. SE-BasicBlock and SE-Bottleneck are obtained by embedding the SE block into the residual modules.

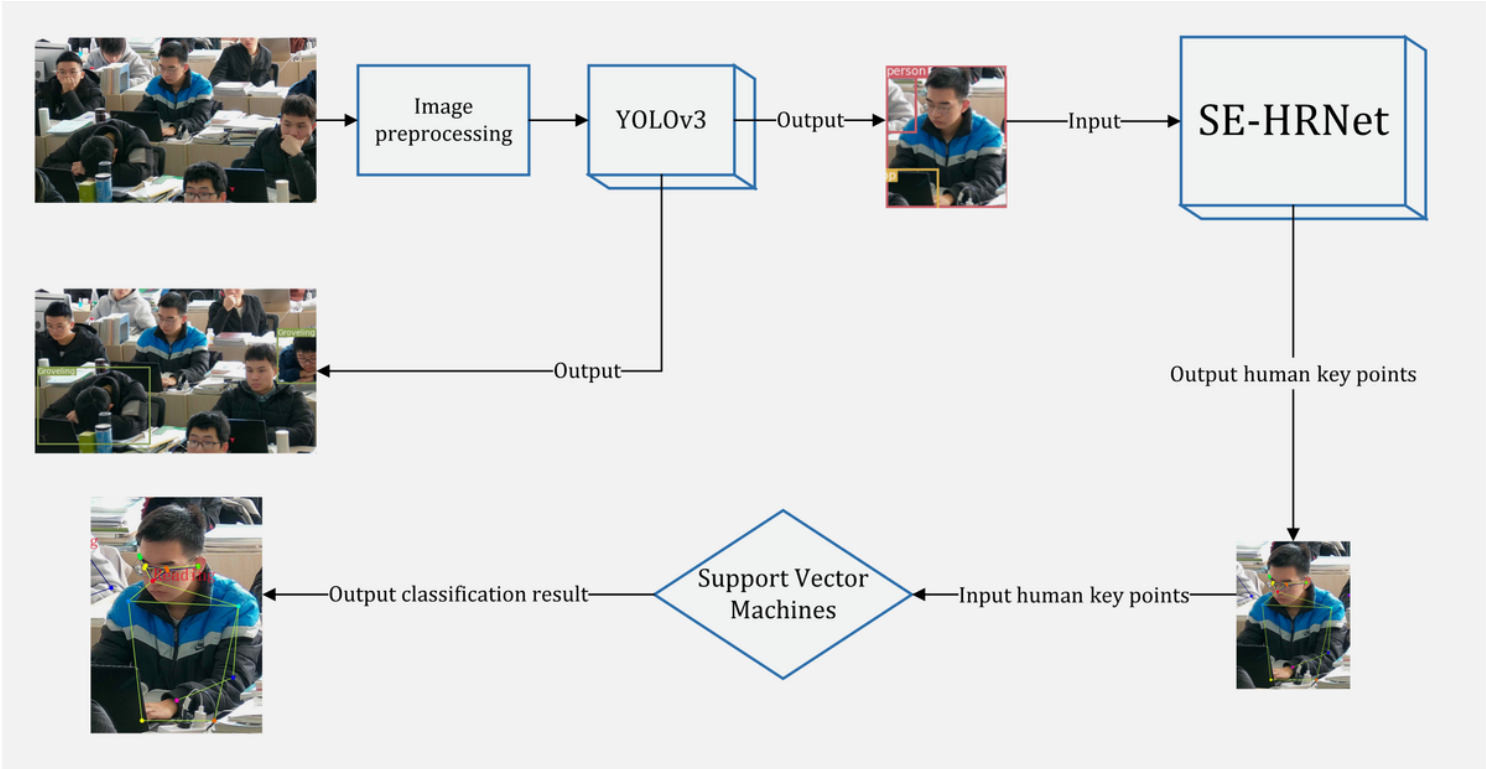

Figure 3

The overview of our method First, we use pre-trained YOLOv3[9] to detect the images we collect from classrooms. The output results fall into two categories, one is the human body object provided for SE-HRNet pose estimation, and the other is the Groveling pose object. Then the results of the Groveling pose object are directly output, and the results of the human body object are cropped from the image. The cropped images are input into SE-HRNet for pose estimation. SE-HRNet detects the locations of 17 keypoints of the human body. The next step is to preprocess the data of output keypoints. We design an SVM classifier to classify the preprocessed keypoints of the human body. Then output the classification results. Finally, the proposed method applied to online de-tection of real surveillance images of the classroom.

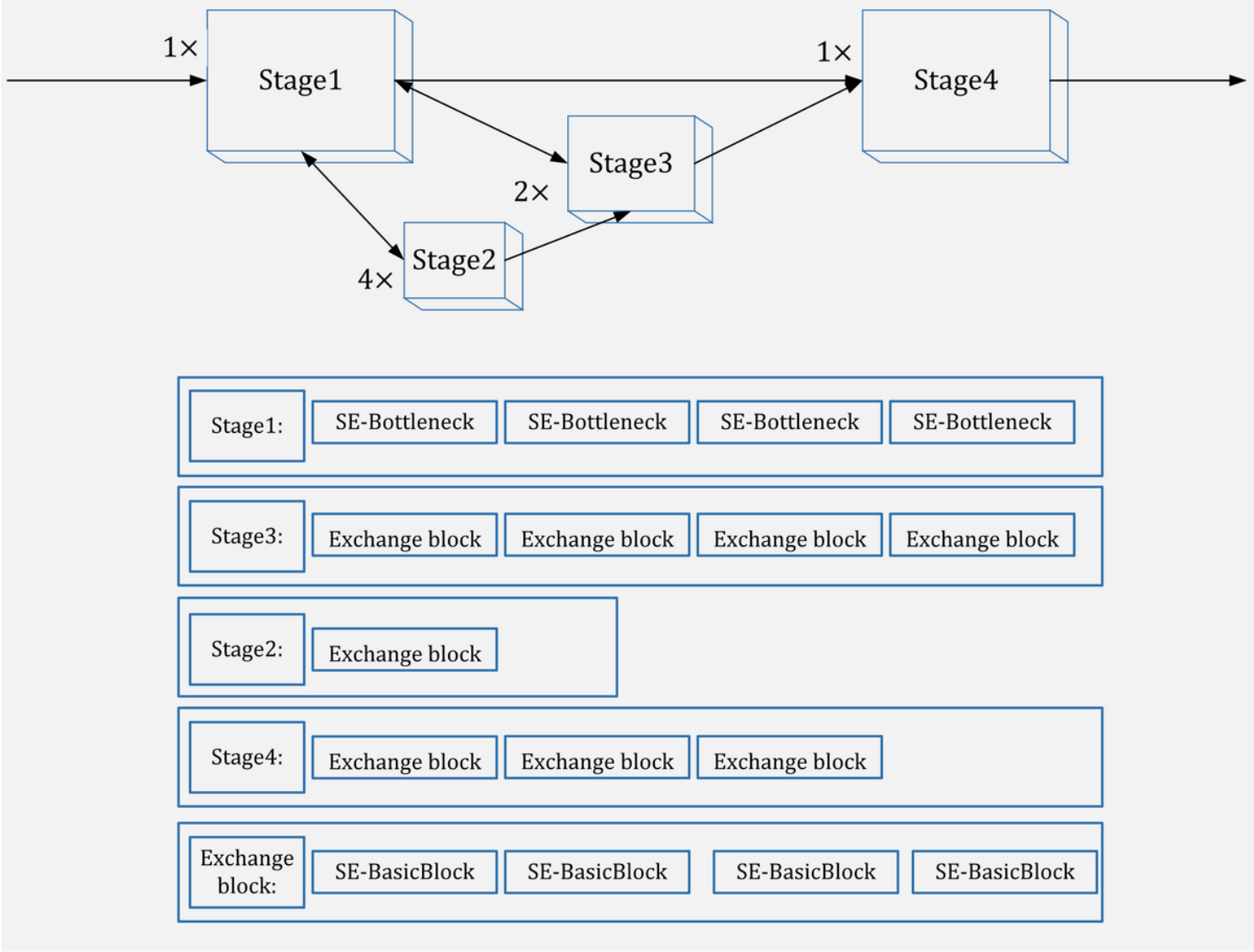

Figure 4

The architecture of the SE-HRNet SE-HRNet consists of four stages with four parallel subnetworks, the resolution is gradually reduced to half and the width (number of channels) is correspondingly in-creased twice.

Figure 5

The type of poses annotated in our dataset To identify students' learning status during class. We labeled a total of three class-room postures of the students, Reading(lift) Groveling(middle) Looking(right).

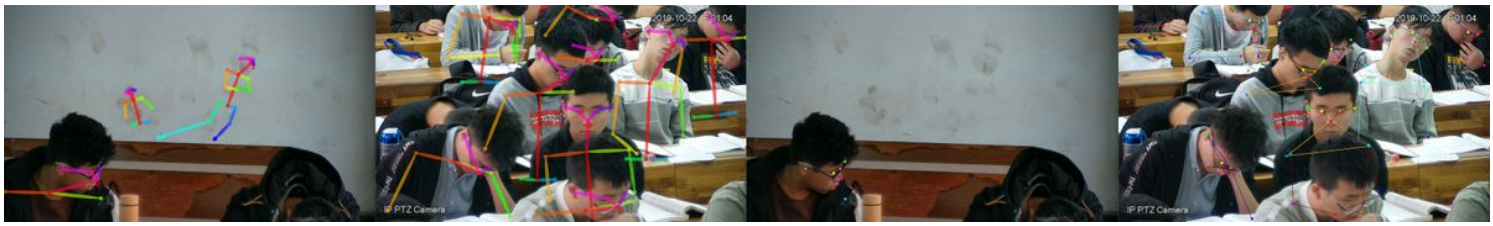

Figure 6

Comparison of the pose estimation results The pose estimation results of our method (right two images) and OpenPose method (left two images). OpenPose mistakenly identified the patterns of the wall as human bodies and the human pose estimation accuracy is not good. SE-HRNet compares to OpenPose obtains significant improvements: reduced the human body object detectd error rate and increased the accuracy of estimate keypoints.

Figure 7

The results of our proposed method. It shows the representative recognized results of our proposed method of sparse to dense situations. In which each pose is labeled right next to body keypoints, and groveling pose is labeled the bounding-box in the images.