Zhang *et al. J Wireless Com Network* (2021) 2021:140
https://doi.org/10.1186/s13638-021-02015-0

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Classroom student posture recognition based on an improved high-resolution network

Yiwen Zhang[1], Tao Zhu[1,3], Huansheng Ning[2] and Zhenyu Liu[1,3*]

*Correspondence:
lzy@usc.edu.cn
[1] School of Computer, University of South China, Hengyang 421001, Hunan, China
Full list of author information is available at the end of the article

Abstract

Due to the large number of students in a typical classroom and crowded seating, most features of student posture are often obscured, making it difficult to balance the accuracy in identifying student postures with computational efficiency. To solve this issue, a novel classroom student posture recognition method is proposed. First, to recognize the poses of multiple students in the classroom, we use the you-only-look-once (YOLOv3) algorithm for object detection and retrain it to detect human objects that are hunching on a table, creating the pose estimation network. Next, to improve the accuracy of the pose estimation network, we use the squeeze-and-excitation network structure that is embedded in the residual structure of high-resolution networks (HRNet). Finally, with the improved HRNet algorithm's outputs of key human body points, we design a pose classification algorithm based on a support vector machine, to classify human poses in the classroom. Experiments show that the improved HRNet multi-person pose estimation algorithm yields the best mean average precision performance of 73.76% on the common objects in context (COCO) validation dataset. We further test the proposed algorithm on a customer dataset collected in a classroom and achieved a high recognition rate of 90.1% and good robustness.

Keywords: Pose estimation, Support vector machine, High-resolution networks, Squeeze-and-excitation networks, Object detection

1 Introduction

In recent years, due to the growth of surveillance systems for both public and personal usage, pose estimation and detection methods have been developed to meet the emerging needs of various industries. For example, there are many abnormal behaviors that university students may exercise in the classroom, such as sleeping, playing on mobile phones and chatting, which can greatly affect students' learning and academic performance in the long term. Therefore, in the scenario of a smart university classroom, the task of student pose estimation and detection using computer vision technology has important research implication and great application value.

There are two categories of solutions for this task. One category is based on object detection algorithms. Lin Tang and Bin T's method [1, 2] uses the improved Faster R-CNN [3] model for object recognition to detect student postures in classrooms. Li W's method [4] only detects students sleeping based on improved R-FCN [5]. These methods

can detect poses in small, low-quality pictures of a classroom where students are concentrated and the collected pictures have low resolution. However, using object detection methods to estimate the poses of each individual on this occasion is hindered by two primary obstacles. First, if we need to detect a new pose, the entire network in the model must be retrained. Second, each method only identifies poses that significantly differ from others, such as sitting and standing; other less distinct poses, such as reading, chatting and raising hands, are typically not recognized.

The other category is based on the pose estimation network. Zaletelj's method [6] employs OpenPose [7] to estimate the location of key points on the human body and then uses a classifier to classify the collected key points. The advantage of this method is that it is easy to use and has a fast calculation speed; however, it also suffers from the low accuracy. Liu's method [8] uses pose estimation maps (heatmaps), the byproduct of pose estimation, to recognize human action. This method is effective when applied on behaviors that have large movements; however, it cannot properly identify behaviors with small movements.

Due to the large number of students in classrooms and bodies being covered by objects such as tables or other bodies, there are four major challenges in recognizing students' postures in a classroom.

1. The estimation of human-body key points has a high error rate and low accuracy rate.
2. Some human joints are invisible to cameras due to occlusions, resulting in only a few unreliable features to estimate human body key points. Therefore, it is difficult to recognize the hunched posture.
3. Most top-down pose estimation methods have low calculation speeds; thus, the final results cannot be made available in real time.
4. If only object detection is used to recognize postures in the classroom, the model may only be able to detect a single gesture with poor scalability.

To tackle these difficulties, we propose a posture recognition method for use in a classroom that combines the pose estimation algorithm and the object detection algorithm.

The contributions of this paper are threefold. First, we capitalize on the you-only-look-once (YOLOv3) model [9] to detect human objects and students hunching on tables. Second, we propose an improved HRNet model as the pose estimation algorithm to reduce the error rate of estimating human body key points. We term our proposed pose estimation algorithm as SE-HRNet, which is constructed by embedding the SENet [10] structure into the HRNet [11]. Finally, we design a posture classification network based on the support vector machine (SVM) [12]. Experimental results show that the mean average precision (mAP) of using YOLOv3 to detect the hunching pose is 91.6%; the mAP of using the SE-HRNet model to detect key points of the human body is 73.7%; the accuracy of the pose classification is 88.6%; Meanwhile the computation speed of the proposed classroom student posture recognition method is 7 images per second in our experimental setting. The remainder of this paper is organized as follows. Section 2, gives a systematic review of related work. The proposed classroom student postures recognition method is detailed in Sect. 3. Experimental details including the proposed

dataset and the key results are presented in Sect. 4. This paper is concluded in Sect. 5 with some discussion on potential future work.

2 Related works

2.1 Pose estimation methods

Currently, multi-person human pose estimation methods can be divided into two categories:

2.1.1 Top-down approaches

First, object detection is performed on human bodies in an image, each human body is cropped into single images. Then, single-person pose estimation is used for each cropped human body. Therefore, for each detection, a single-person pose estimator is run, and the more people there are, the greater the computational cost. However, the accuracy of the top-down method is typically higher; the common models include CPN [13], hourglass [14], CPM [15], alpha pose [16], etc.

Bottom-up approaches: First, the model detects all key points of the human body in the picture then matches these points to different individuals; thus, this method is faster in calculation but yields marginally lower accuracy than that of the top-down method. The most common bottom-up model is OpenPose [7].

The high-resolution network (HRNet) [11] is a human body pose estimation method that is an example of a top-down method. High-resolution network pose estimation can maintain high-resolution representations through the whole process. It begins from a high-resolution subnetwork in its first stage, gradually adding high-to-low resolution subnetworks one by one to form more stages and connecting the multiresolution subnetworks in parallel [11]. The network then performs multiple multiscale fusions by repeatedly exchanging information across parallel multiresolution subnetworks and estimates the key points of the human body via the high-resolution representations of the network output. The architecture of HRNet is shown in Fig. 1.

HRNet has two benefits compared 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 networks do. Therefore, HRNet can maintain high resolution rather than restore resolution using a low- to high-resolution process. Thus, the predicted heatmap is spatially more precise. Second, most existing fusion schemes combine low- and high-level representations [11]. Conversely, this method uses the low-resolution representation of the same depth and a similar level to perform multiple multiscale fusions to improve the high-resolution representation, and vice versa, giving the high-resolution representation detailed pose estimation data. Thus, this method yields more accurate heatmaps.

HRNet can maintain high-resolution features without the need to recover the high resolution. HRNet also fuses parallel multiresolution representations repeatedly, enhancing the reliability of high-resolution representations, yielding accurate and spatially precise point heatmaps. However, because HRNet is a top-down method, its image processing speed is typically slower than that of a bottom-up method. Additionally, to achieve HRNet multi-person pose estimation, the object detection algorithm must process the

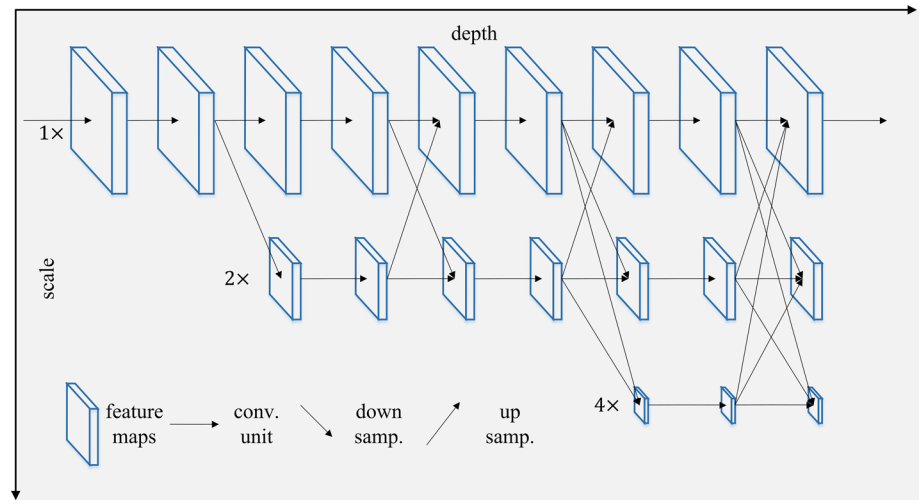

Fig. 1 Architecture of the HRNet. The architecture of the HRNet. It consists of parallel high-resolution and low-resolution subnetworks with repeated information exchange between multiresolution subnetworks (multiscale fusion)

image first. Therefore, the detection speed of the object detection algorithm has a strong impact on the pose estimation speed.

2.2 Squeeze-and-excitation networks

Squeeze-and-Excitation Networks (SENet) [10] introduced a new architectural unit called squeeze-and-excitation (SE) blocks with the goal of improving the quality of representations produced by a network by explicitly modeling the interdependencies between the channels of conventional features [10]. In this structure, squeeze and excitation are two critical operations. A new "feature recalibration" strategy is used, through which 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 Fig. 2, where SE represents the SENet blocks. The first block passes through a squeeze operation, which first performs global average pooling on the input feature map to obtain a feature map of size $C \times 1 \times 1$, where 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. Aggregation is followed by an excitation operation, through which the parameter W is used to generate weights for each feature channel, where the parameter W is learned from the correlation between the feature channels. After two fully connected layers (first dimensionality reduction and then dimensionality increase), the method uses the sigmoid activation function to obtain a weight of $C \times 1 \times 1$, followed by a re-weighting 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 via multiplication to complete the feature recalibration. The output of the SE blocks can be fed directly into subsequent layers of the network. Both BasicBlock and Bottleneck are the classic residual modules used in ResNet. SE-BasicBlock embeds the SE structure into the regular BasicBlock unit, SE-Bottleneck embeds the SE structure into the regular Bottleneck unit.

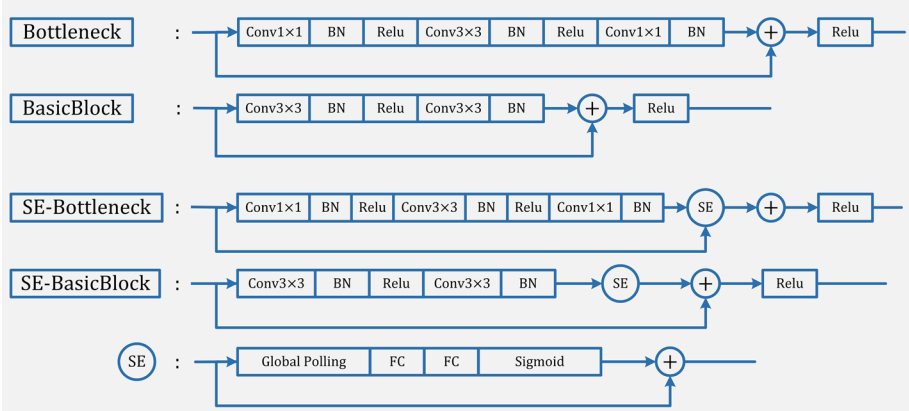

Fig. 2 Architecture of SENet. BasicBlock and Bottleneck are original Residual modules. SE-BasicBlock and SE-Bottleneck is obtained by embedding the SE block into the residual modules

The structure of the SE block is simple, so it can be directly embedded into existing network architectures, which markedly improves results, is computationally lightweight, and imposes only a marginal increase in model complexity and computational burden.

2.3 Object detection method

Existing object detection algorithms are primarily divided into two types: the two-stage method (i.e., the region proposal method) and the one-stage method (i.e., the regression method). Two-stage object detection algorithms include RCNN [17], Fast RCNN [18] and Faster-RCNN [3]. Faster R-CNN tends to be a slower but more accurate model [19] and 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 proposal boxes are used to crop features from the same intermediate feature maps, which are then entered into the feature extractor to predict a class for each proposal box and to optimize each proposal box.

In a one-stage object detection algorithm (e.g., SSD [20] and YOLO [21]), object classification and bounding-box regression are conducted concurrently without a region proposal stage [22]. YOLO converts object detection into regression work. Based on a single end-to-end network, calculations are completed from the original image to the output of the object position and category. These one-stage methods typically exhibit a high detection speed and high efficiency but low accuracy. YOLOv3 [9] can detect multiple objects with a single inference; thus, its detection speed is high. Additionally, using a multistage detection method, YOLOv3 improves upon the low accuracies of YOLO and YOLOv2 [23]. However, YOLOv3 yields lower detection accuracy than Faster-RCNN with small targets. However, the detection speed of YOLOv3 is marked higher than Faster-RCNN [3]. Therefore, YOLOv3 is suitable for many engineering applications.

Considering the detection speed and accuracy of the algorithm, this paper uses YOLOv3 for object detection to detect the human body and hunched postures in the classroom, and provides the foundation for the proposed real-time classroom human posture recognition method based on SE-HRNet.

3 Methods

3.1 Overview of the framework

An overview of the classroom student postures recognition method proposed in this paper is shown in Fig. 3. First, we use pretrained YOLOv3 [9] to detect images collected from classrooms. Results fall into two categories: one is the human body object provided for SE-HRNet pose estimation, and the other is the hunched posture object. Then, the results of the hunched posture 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, which detects the locations of 17 key points of the human body. The next step is to preprocess the output key points. We thus design an SVM classifier to classify the preprocessed key points of the human body, and then output the classification results. Finally, the proposed method is used for real-time posture recognition of student images in a classroom.

3.2 YOLOv3 application

Due to the limitations of classroom usage scenarios, the results of student posture recognition must be shown in real time and must be as accurate as possible. Therefore, we must address 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 [3]. Based on the discussion in Sect. 2.3 of this paper, we propose replacing Faster R-CNN with YOLOV3 [9] for object detection in the proposed method.

Among the three poses we proposed to recognize, the hunched posture is the most difficult to recognize using a pose estimation network. Because the hunched posture

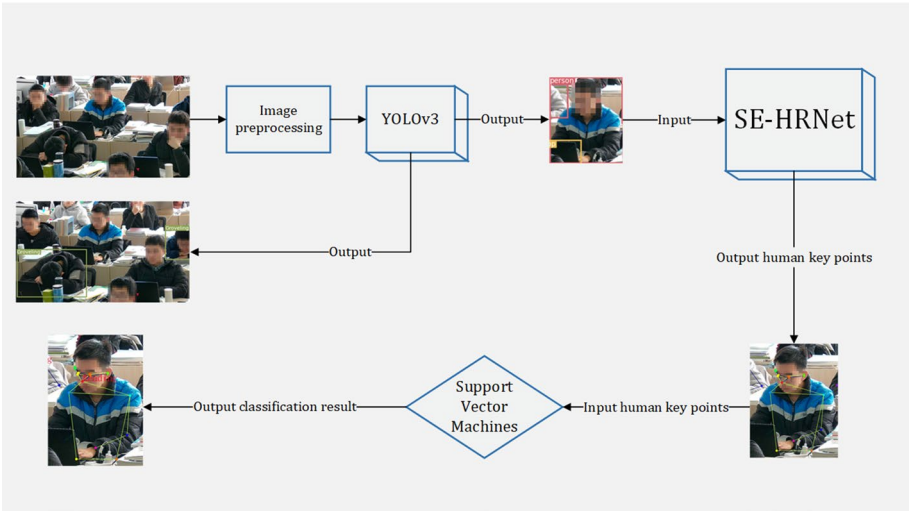

Fig. 3 Overview of our method. First, we use pretrained YOLOv3 [9] to detect the images we collected from classrooms. Results fall into two categories: one is the human body object provided for SE-HRNet pose estimation, and the other is the hunched posture object. Then the results of the hunched posture 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 key points of the human body. The next step is to preprocess the data of output key points. We design an SVM classifier to classify the preprocessed key points of the human body. Then output the classification results. Finally, the proposed method applied to online detection of real surveillance images of the classroom

means that the person is hunched over the table, usually only the top of his head is shown in the photo, key points of human body will be seriously lost if we try to use the pose estimation network to estimate the hunched posture. This fact makes pose estimation of hunched posture impossible.

To solve this problem, we use an object detection network to detect the hunched posture. We also need use an object detection network to detect human objects in the classroom for the pose estimation network. Therefore, we use the datasets we collected to retrain YOLOv3 to detect the hunched posture and improve the accuracy of human object detection in the classroom.

The proposed method uses YOLOv3 to detect the hunched posture and to improve the existing OpenPose method, which cannot estimate the hunched posture and has difficult in recognizing heavily occluded human body key point postures.

3.3 Designing the improved HRNet

When human bodies overlap, many human body features are occluded; this is particularly true in crowded and complex places, such as classrooms. Conventional pose estimation networks output feature maps that have high confidence in the key points of the overlapping parts. The network mistakenly believes that the overlaps or missing key points are also part of the human body. This unbalanced confidence distribution causes many misidentifications [24]. To enable the network to learn more global features, enhancing the receptive field can be used to balance the confidence of the heat map in different positions. Therefore, we propose embedding 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 fully connected layers serve to reduce the number of parameters. HRNET for feature extraction is the key to accurately estimating the key points, which the residual layer fuses into multiple layer features. Therefore, we propose embedding the SENet structure into BasicBlock and Bottleneck of the HRNet to obtain the SE-BasicBlock and SE-Bottleneck substructures (see Fig. 2), thereby expanding the receptive range of the feature map to global information.

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

The SE structures introduce primitive information into deep layers, inhibit information degradation, expand the receptive field by pooling, and then integrate shallow information with deep information from multiple dimensions so that the combined output contains multiple levels of information, enhancing the feature map's expression ability.

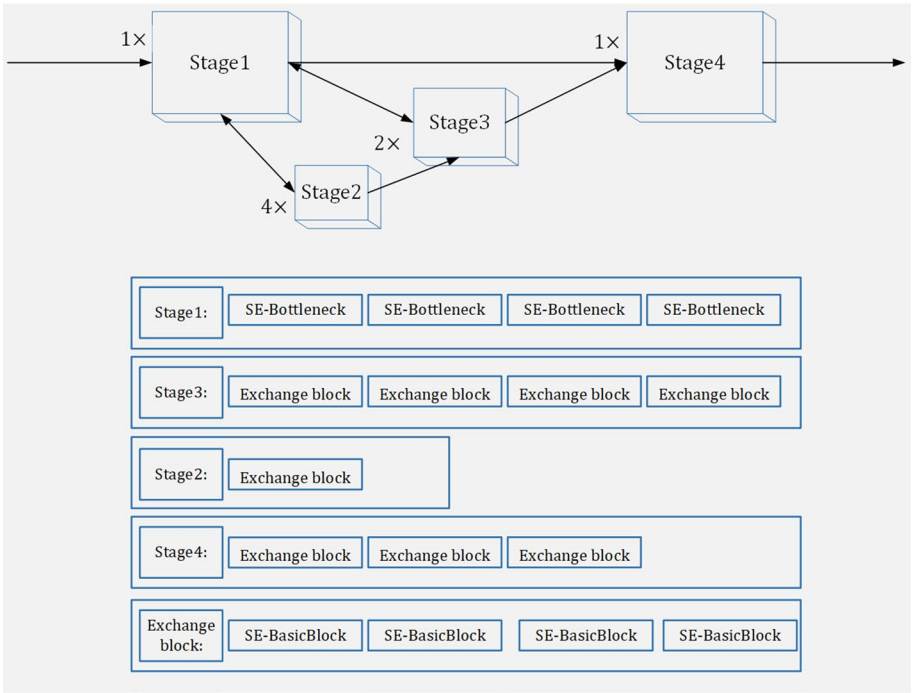

Fig. 4 Architecture of the SE-HRNet. SE-HRNet consists of four stages with four parallel subnetworks: the resolution is gradually reduced to half, and the width (i.e., number of channels) is correspondingly increased twice

Fig. 5 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 wall patterns as human bodies, and the human pose estimation accuracy was poor. SE-HRNet yields significant improvements over OpenPose, reducing the human-body-object-detected error rate and increasing the accuracy of estimate key points

The results of the SE-HRNet tested on the proposed dataset are compared to OpenPose, as shown in Fig. 5. OpenPose mistakenly identified wall patterns as human bodies, and the human pose estimation accuracy was poor. SE-HRNet yields significant improvements over OpenPose, reducing the human-body-object-detected error rate and increasing the accuracy of estimated key points.

The experimental results show that the detection accuracy of HRNet is significantly enhanced by introducing the SE structure, which reduces high detection error rates compared to existing methods that use OpenPose.

3.4 Designing the classification method

3.4.1 Data preprocessing

To reduce the number of calculations, speed up convergence, and improve accuracy, the human body key point data output from SE-HRNet must be preprocessed. First, because the coordinate origin of the key point data output by SE-HRNet is in the upper left corner of the image, and each image contains multiple human bodies, it is necessary to shift the coordinate origin to the nose position of the 17 points in each human body. Then, we normalize the data and scale the coordinate data to between 0 and 1 based on the image resolution.

3.4.2 Designing the SVM classifier

The classifier structure is a simple four-layer fully connected network. Each layer has 125 neurons and uses a rectified linear unit (ReLU) as the activation function. We use the Adam optimizer, the base learning rate is set as 1e-3, and the training process is terminated after 150 epochs. We only use a classifier to classify two types of actions in the classroom: reading and looking. The loss function used is the hinge loss form of the support vector machine (SVM; see Eq. (4)) [12]. The simplest way to extend SVMs for multiclass problems is using the so-called one-vs-rest approach [25]:

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

The classroom student posture recognition method combines object detection, pose estimation and key point classification. Therefore, if we want to recognize a new pose in the classroom, we only retrain the key point classification network instead of retraining all networks in the method. Therefore, the proposed method improves scalability compared to existing methods by combining three different models.

4 Experimental

4.1 Dataset and Parameter Settings

4.1.1 Dataset

In this experiment, the COCO2017 dataset is used to train and validate the improved HRNet [26]. This dataset includes 149,808 pictures and over 250,000 person instances labeled with 17 key points. We evaluate the improved HRNet with the val2017 data set, which contains 6384 pictures.

To describe the situation in the real classroom environment as much as possible, we collected a dataset that included pictures of students in several classrooms during class. This dataset contains many images with different degrees of occlusion and light changes (see Fig. 5). This dataset was created using Dahua network dome surveillance cameras. Two dome cameras were installed in a large classroom of 120 people, and one was installed in a small classroom. A total of 10 cameras were installed in 4 large rooms and 2 small rooms in 6 classrooms of the university. If we take a picture of the entire classroom directly, details are lost; however, we also cannot take a picture of everyone in the classroom. Therefore, we used a sampling method. First, we set a fixed number of cruising

Fig. 6 Detection results of our proposed method. It shows the representative recognized results of the proposed method of sparse to dense situations. Each pose is labeled immediately next to body key points, and hunched posture is identified the bounding box in the images

points (30 points for a camera in a large classroom and 20 points for a small classroom), and the photos collected from these points covered every seat in the classroom. During

Fig. 7 Type of poses annotated in the proposed dataset. To identify students' learning status during class. We labeled a total of three classroom postures of the students, reading(lift) hunching(middle) looking(right)

a class, the camera moves to a cruising point every 15 s to collect two photos at a resolution of 2592×1520. The advantage of this method is that it can cover every seat in the class, and the collected high-resolution data can be kept intact. The disadvantage of this method is that each collected picture inevitably contains incomplete human body data. As shown in Fig. 6. Detection results of the proposed method. Figure 6, there is always a certain number of incomplete human body data around the upper, lower, left and right sides of the picture. We collected more than 40,000 images; removed many poor-quality images and images with no information about the human body; and ensured that the training and test sets received the same proportion of categories as the dataset. The data collected in this paper contained 1951 training samples and 943 test samples. Each picture has an average of 5 people, and we labeled a total of 14,470 student body postures. There are three types of poses annotated in this dataset, including reading, hunching and looking, as shown in Fig. 7. The ratio of these three categories in the data set is 5:4:1. In a classroom environment, there are relatively few hunched postures.

4.1.2 Experimental environment

The software environment used in this study included Ubuntu 18.04 based on PyTorch 1.4 with CUDA 10.1. The hardware environment included an Intel Core i7 7820X CPU, 64 GB RAM, and an Nvidia TITAN X (Pascal) 12G graphics card.

Table 1 Comparison of different methods using the COCO validation set

Method	Input size	#Params	GFLOPs	AP	AR
OpenPose [7]	368 × 368	–	–	61.8	66.5
Baseline ResNet-50 [28]	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

4.1.3 Evaluation metrics

The standard evaluation metric of the pose estimation experiment is based on object key point similarity (OKS) in Eq. (5), where d_i is the Euclidean distance between the detected key points and the corresponding ground truth; v_i is the visibility flag of the ground truth; s is the object scale; and k_i is a per-key-point constant that controls falloff. We report standard average precision and recall scores [11], where 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 applies to the average recall (AR).

The evaluation measurements of the object detection algorithm are the average precision (AP), which was proposed in [27]. A common judgment for the correctness is the intersection-over-union (IOU) between the 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]. To obtain a higher recall, we set the IOU threshold to 0.5:

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

5 Results

5.1 Training YOLOv3 on the proposed dataset

YOLOv3 uses the Darknet-53 backbone and pretrains the backbone on the COCO dataset. We set the network input resolution to 416 × 416 and use multiple scale training. The dataset we use to retrain and test YOLOv3 is the dataset containing real classroom images that we collected. We retrain YOLOv3 to detect the hunched posture and student bodies, and the training process is finished within 150 epochs. After retraining YOLOv3, the best result on the test set for the hunched posture is AP = 91.6%.

5.2 Comparing different pose estimation methods

To verify the effectiveness of the improved HRNet, several pose estimation frameworks are investigated, including OpenPose [7], original HRNet [11], and ResNet [28], for comparison.

Both the original HRNet [11] and SE-HRNet were trained on COCO2017 [26] with an input size of 256 × 192. The learning and dropout rates remain unchanged based on the settings in [11], and the training is set for a total of 210 epochs. HRNet has one small net and one big net: HRNET-W32 and HRNET-W48.Where 32 and 48 represent the widths of the high-resolution subnetworks in last three stages, respectively. The smaller net HRNET-W32 was used in our experiment.

Table 1 shows the results of the proposed improved HRNet compared to other multi-person pose estimation methods on the COCO verification set. The improved HRNet with an embedded SENet structure achieves an AP score of 73.7, outperforming other methods with the same input size (256×192) except OpenPose [7]. The OpenPose uses an input size of 368×368 and is a bottom-up approach. The proposed approach yields much more accurate results than the bottom-up approach: the proposed improved network improves AP by 11.9% compared to OpenPose. Compared to SimpleBaseline-ResNet-50 [28], the proposed model yields marked improvements: a gain of 3.3% with a smaller model size and fewer GFLOPs.

Based on these results, the proposed HRNet [11] training results exhibited a marginal decrease (0.3%) compared to the training results provided in [11]. HRNet's GFLOPs and number of parameters were also marginally higher. The results of the SE-HRNet compared to those of the original HRNet yielded 0.7% of improvement, and its model size (#Params) and GLOPs did not increase significantly (approximately 1%).

5.3 Comparing 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 posture recognition method. However, OpenPose's pose estimation is not sufficiently accurate and yielded many classification errors, preventing the hunched posture from being recognized. Because the human body features of the hunched posture are frequently occluded, OpenPose cannot output any useful human body key points.

Second, the method using Faster RCNN + HRNet + SVM exhibited certain improvements compared to the method using OpenPose and could recognize the hunched posture because Faster RCNN was used to detect the hunched posture. Additionally, the accuracy of Faster RCNN in detecting the hunched posture was high because Faster RCNN is a two-stage object detection algorithm.

Finally, the YOLOv3 + SE-HRNet + SVM proposed in this paper yielded significant improvements (8.3%) compared to other methods, reaching 90.1% accuracy. Although YOLOv3 is a one-stage object detection algorithm, the accuracy of detecting the hunched posture is similar to Faster RCNN.

5.4 Computational costs

To evaluate computational costs, we tested different approaches on a PC with the same configuration as described above. These results are shown in Table 3, where no method includes the last step of pose classification.

Table 2 Comparisons of different methods on the proposed dataset

Method	Reading (%)	Looking (%)	Hunching (%)	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

Table 3 Comparison of the processing time of different methods on the proposed 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

The average running time of each image in the proposed method is 0.142 s. Thus, the proposed method is marginally slower than OpenPose because OpenPose is a bottom-up method. However, the proposed method yields a similar time to the HRNet + YOLOV3 method and a significantly faster time than the HRNet-W32 + Faster RCNN method by 404%. These results also show that the proposed improved HRNet does not add much computational cost based on its addition of the SENet structure, indicating that the proposed approach works well in real classroom environments.

6 Discussion

This paper introduces a new approach to multi-student posture recognition in a classroom environment based on an improved high-resolution network. Specifically, this method combines the YOLOv3 object detection algorithm and the HRNet pose estimation network and further enhances HRNet with the SENet structure, leading to SVM-based pose classification algorithms. The proposed approach and methods have been tested and evaluated using the COCO validation dataset and a customer dataset, and both yield impressive results. Most existing classroom posture recognition methods are object detection or pose estimation methods; however, the proposed method combines object detection, pose estimation and neural network classification algorithms for multi-student posture recognition. The method proposed in this paper is also compared to the methods used in other papers [1, 2, 4], which use object detection technology for multi-student posture recognition in a classroom environment. The proposed method can recognize more postures and exhibits better scalability and robustness. Compared to J. Zaletelj's paper [6], which uses the multi-person pose estimation method, the proposed method improves HRNet using the SENet structure and yields high-accuracy pose estimation in complex cluttered classroom environments.

Figure 5 shows the representative recognized results of the proposed method from sparse to dense situations. Each pose is labeled immediately next to body key points, and the hunched posture is identified by a bounding box in the images. The proposed method can manage sparse and concentrated distributions of students, and can clearly locate students and recognize their poses in difficult situations, even with occlusions. In another difficult situation, where certain hunch postures can easily be confused with the looking pose, the proposed method still frequently predicts the correct label. In other cases, such as with occlusions or background noise, the proposed method also exhibited more robust results than other methods.

Certain limitations in this study do exist. Due to limited manpower, we collected a small amount of data in the customer dataset used in this study. The amount of data

used to test the proposed methodology is not large and thus cannot represent all classroom environments.

7 Conclusions

In this paper, we propose a classroom student posture recognition method that effectively combines pose estimation, object detection and posture classification with strong robustness and scalability. We choose YOLOv3 as the object detection network to detect hunch postures due to its efficiency. Then, to alleviate the high error rates of other common pose estimation methods, we propose to embed the SENet structures into HRNet. Experiments on the COCO dataset show that the AP of the improved HRNet reaches 73.8%, slightly higher than the original HRNet. Finally, we design a posture classification algorithm based on the SVM and the accuracy of the proposed method reaches 90.1%, outperforming other traditional methods. In the future work, we will adapt the posture classification algorithm to recognize more types of student postures. Another extension would be to acquire extra student data in different environments to further improve the generalization ability of the proposed approach.

Abbreviations

HRNet: High-resolution network; SENet: Squeeze-and-excitation networks; YOLO: You-only-look-once; SVM: Support vector machine; COCO: Common objects in context.

Acknowledgements

Not applicable.

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.

Funding

This work is partly supported by the National Natural Science Founda-tion of China (No. 61872038, 62006110), Natural Science Foundation of Hunan Province (No. 2019JJ50499).

Availability of data and materials

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

Declarations

Competing interests

The authors declare that there is no conflict of interest.

Author details

[1]School of Computer, University of South China, Hengyang 421001, Hunan, China. [2]School of Computer and Communication Engineering, University of Science & Technology Beijing, Beijing, China. [3]Hunan Provincial Base for Scientific and Technological Innovation Cooperation for Medical Big Data, Hengyang, China.

Received: 6 September 2020 Accepted: 10 June 2021
Published online: 26 June 2021

References

1. L. Tang, C. Gao, X. Chen, Pose detection in complex classroom environment based on improved Faster R-CNN. IET Image Proc. **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. Mod. Comput. (2018)
3. S. Ren, K. He, R. Girshick, Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **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. Adv. Neural Inform. Process. Syst. 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. Z. Cao, T. Simon, S.-E. Wei, Y. 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. M. Liu, J. 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. J. Hu, L. Shen, G. 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. K. Sun, B. Xiao, 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, in *Computer Vision–ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, vol. 9912, ed. by B. Leibe, J. Matas, N. Sebe, M. Welling (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. 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. J.R.R. Uijlings, K.E.A.V.D. Sande, T. Gevers, Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
18. R. Girshick, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015 IEEE Conference. Fast R-CNN (Santiago, Chile, 2015), pp. 1440–1448
19. J. Huang, V. Rathod, C. 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
20. W. Liu et al., SSD: single shot multibox detector, in *Computer Vision–ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, vol. 9905, ed. by B. Leibe, J. Matas, N. Sebe, M. Welling (Springer, Cham, 2016)
21. 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
22. J. Choi, D. Chun, H. 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
23. 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
24. X. Liu, Li. Yuqian, L. Li, Improved YOLOV3 object recognition algorithm with embedded SENet structure. Comput. Eng. **45**(11), 243–248 (2019)
25. Y. Tang, Deep Learning using Linear Support Vector Machines. (arXiv, 2013), https://arxiv.org/abs/1306.0239. Accessed 2 June 2013
26. T. Lin, M. Maire, S.J. Belongie, Microsoft COCO: common objects in context (2014), https://cocodataset.org. Accessed 2014
27. R. Padilla, S.L. Netto, 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
28. B. Xiao, H. Wu, Y. Wei, in 15th European Conference on Computer Vision (ECCV). Simple Baselines for Human Pose Estimation and Tracking (Munich, Germany, 2018), pp. 472–487

Publisher's Note