INTELLIGENT STUDENT BEHAVIOR ANALYSIS SYSTEM FOR REAL CLASSROOMS

Rui Zheng *Fei Jiang*⋆ *Ruimin Shen*

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
zhengr@sjtu.edu.cn, jiangf@sjtu.edu.cn, rmshen@sjtu.edu.cn

ABSTRACT

In this paper, we design an intelligent student behavior analysis system for recorded classrooms, which automatically detects hand-raising, standing, and sleeping behaviors of students. Detecting these behaviors is quite challenging mainly due to various scale behaviors, low resolution, and imbalanced behavior samples. To overcome the above-mentioned challenges, we first build a large-scale student behavior corpus from thirty schools, labeling these behaviors using bounding boxes frame-by-frame, which changes the behavior recognition problem into object detections. Then, we propose an improved Faster R-CNN, a classical object detection model, for student behavior analysis. Specifically, we first present a novel scale-aware detection head to overcome scale variations. Secondly, we propose a new feature fusion strategy to detect low-resolution behaviors while introduces little computation overhead. Thirdly, we utilize OHEM (Online Hard Example Mining) to alleviate severe class imbalances. Experiment results on our real corpus are increased by 3.4% mAP while maintaining a fast speed.

Index Terms— student behavior detection, Faster R-CNN, scale-aware detection head, feature fusion

1. INTRODUCTION

Student behavior analysis in classrooms is not only an essential part of evaluating the quality of teaching, but also helpful to predict student long-term development [1, 2]. However, previous student behavior analysis largely depends on the observations of teachers, which is time-consuming and can not fulfill large-scale practical requirements. Therefore, an automatic student behavior detection is needed for comprehensive and in-depth analysis. In this paper, we focus on designing an intelligent student behavior system.

Existing behavior detection algorithms can be roughly divided into three categories: hand-crafted feature-based, pose-estimation-based and object-detection-based. The hand-crafted feature-based algorithms like [3] and [4] combined hand-crafted shape and motion features to describe the human behavior information, which highly rely on the environments, and perform poorly on the complex real classroom scenarios; The pose-estimation-based algorithms [5, 6] applied pose estimation method to generate human joints for behavior detection, which is not applicable due to the overcrowded classrooms in China; The object-detection-based algorithms are recently proposed [7, 8] and received impressive results. In this paper, we also adopt an object-detection-based algorithm for student behavior analysis.

As for object detection, the two-stage R-CNN frameworks [9, 10, 11] have received more attention due to their impressive detection results on public datasets. However, the

Fig. 1: **Student behavior detection.** (a) Example of student behaviors, including hand-raising and standing in a real classroom. (b) Number fraction of objects vs. Relative area to the image in our dataset and MS COCO dataset. (c) Imbalanced category distribution of our dataset.

⋆ Corresponding author.

The work was supported by National Nature Science Foundation of China (No. 61671290), China Postdoctoral Science Foundation (No. 2018M642019), and Shanghai Municipal Commission of Economy and Information (No. 2018-RGZN-02052).

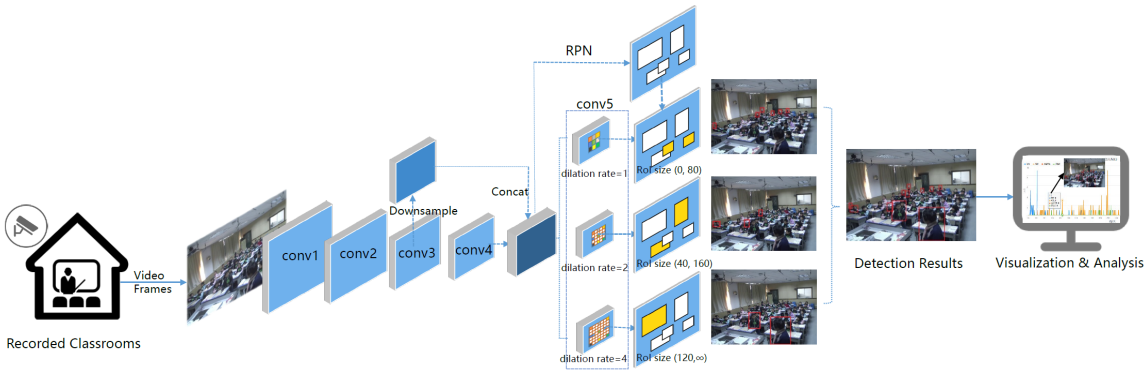


Fig. 2: **The overall architecture of our intelligent student behavior analysis system.**

datasets from real classrooms are quite different from public ones and the classical methods perform poorly in real classrooms. One of the representative issues is large scale variations among different behaviors, such as hand-raising (about 40×40 pixels) and standing (about 200×200 pixels), which results in high scale variations of almost 25 times, as shown in Fig. 1(a). To make matters worse, compared to the most popular object detection dataset MS COCO [12], nearly 70% of the objects in our dataset only occupy less than 0.5% part of the whole image, which introduces the challenge of detecting very small objects in our student behavior datasets. Moreover, our dataset suffers from large class imbalance both between categories and the hard vs. easy examples.

In this paper, an improved Faster R-CNN [11] algorithm is proposed to solve the above-mentioned challenges. First, a scale-aware detection head using different dilation rates is proposed for RoIs with different sizes to solve the scale variation challenge. Second, a simple but efficient feature fusion strategy is proposed to fuse different layers' features for capturing both high-resolution details in low level and semantic information in high level and detecting the rather small objects. Lastly, OHEM (Online Hard Example Mining) [13] is used to focus on harder examples to address the class imbalance in our dataset. We evaluate our methods above both on our student behavior dataset and the selected hard-set from our dataset.

Our main contributions are as follows:

(1) We build a large-scale dataset for student behavior analysis, which contains 70k hand-raising samples, 20k standing samples, and 3k sleeping samples.

(2) We design an intelligent student behavior analysis system for real classrooms using improved object detection algorithms.

(3) Our designed system achieves an increase of 3.4% mAP on our real corpus, which shows the potential for practical applications.

2. RELATED WORKS

In this section, we briefly introduce several existing strategies for scale-invariant detection and class imbalances handling.

2.1. Scale-invariant Detection

The most common way for scale-invariant detection is to exploit pyramid-style methods to enhance the capability of detecting large-scale-variation objects, especially these small objects, such as SSD [14] and FPN [15]. However, these pyramid methods are very time-consuming and memory-costly which are not applicable in real classrooms. Recently, some researches [16, 17] adopted a scale-aware strategy and dilated convolutions [18] for solving scale-variation-detection. However, there are still gaps in small object detection performances compared to the pyramid methods. Inspired by these two methods for scale-invariant detection, in this paper, we first propose a scale-aware detection head which resorts to dilated convolution for different receptive fields. Then, a feature fusion strategy is proposed to capture both low-level and high-level semantic information for small object detection, while avoids heavy computation cost.

2.2. Class Imbalance Handling

In object detection, the common practice to solve class imbalance is to perform some forms of hard examples mining that sample hard examples during training. Online Hard Example Mining (OHEM) [13] proposed a simple but effective way to improve the training of region-based detectors, which has been widely used in object detectors [14, 19]. In this paper, we also adopt OHEM during model training to focus on these hard examples, especially the uncommon sleeping examples and low-resolution hand-raising examples.

3. OUR METHODS

In this section, we first give a detailed introduction to our intelligent student behavior analysis system, including hand-raising, standing and sleeping. Then we elaborate on the strategies we design for alleviating challenges in the real dataset.

3.1. Overall Architecture

The whole system uses video streams from recorded classrooms as the model input. Then our algorithm will output the detection results for further analysis and visualization, as shown in Fig. 2. Specifically, our behavior detection algorithm is based on Faster R-CNN [11]. We adopt ResNet-101 [20] containing 5 blocks as the feature extraction network and these residual blocks are denoted as $\{conv1, conv2, conv3, conv4, conv5\}$ respectively. Different from the original Faster R-CNN, we first use three branches of $conv5$ with dilation rates 1, 2, 4 respectively for RoIs of different sizes. Then, we fuse the output features of $conv3$ with $conv4$ to capture both low-level high-resolution feature maps and high-level semantic information. Lastly, we apply OHEM during training focusing on these hard examples to address the class imbalances.

3.2. Scale-aware Detection Head

In the original Faster R-CNN based on ResNet-101, the $conv5$ layers are deployed as the sub-network (also called detection head) for classification and regression. However, Faster R-CNN attaches RoIs with different sizes to the same detection head and performs poorly on large scale variance. Fig. 1(a) shows extreme scale variance in our student behavior dataset. Scales not only differ between classes such as much larger standing behavior, but also within classes because of large and overcrowded classrooms in China. Inspired by recent papers like SNIP [16] and Trident Network [17], we propose a scale-aware detection head using different dilation rates to detect objects of various sizes.

Specifically, we apply three branches of $conv5$ block with dilation rate 1, 2, 4 respectively based on the feature map from the backbone. The intuition is that we need different receptive fields to detect objects with different sizes. We leverage a scale-aware strategy to assign each branch a specific range of scales matching each receptive field. For RoIs with size range $(0, 80)$, $(40, 160)$ and $(120, \infty)$, corresponding detection head with dilation rates of 1, 2, 4 respectively is performed to capture features at different scales. During training and inference, we apply different detection head to RoIs according to the size range, shown in Fig. 3. Through using dilated convolutions [18], these three branches can share the same network structure and parameters yet having different receptive fields. With this design, our model can detect objects of different scales at different receptive fields without increasing model size. Ablation experiments show the effectiveness of this multi-dilation-branches design.

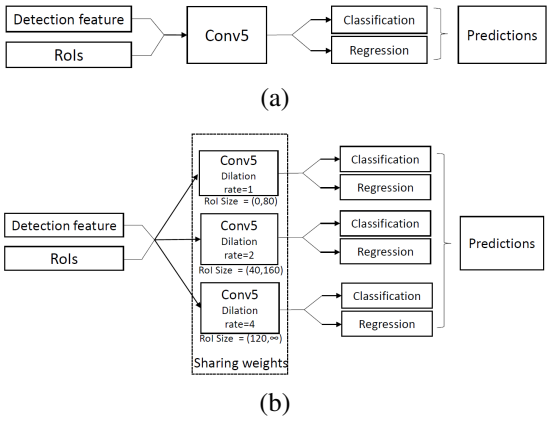


Fig. 3: **The framework of different detection heads.** (a) Original detection head in Faster R-CNN. (b) Our scale-aware detection head using dilated convolution for classification and regression.

3.3. Feature Fusion Strategy

Faster R-CNN with ResNet backbone adopted the output feature map of $conv4$ as the final detection feature. But this low-resolution feature map is not enough for detecting objects smaller than its stride. In real classrooms, most of the behaviors are rather small relative to the original image size. Nearly 70% of the objects in our dataset only occupy less than 0.5% part of the whole image, as shown in Fig. 1(b). Detecting such extremely small objects is challenging, especially for hand-raising with low-resolution and various gestures.

Instead of using costly pyramid-like FPN [15], we propose to fuse the features in a simple but efficient way. Specifically, we combine the outputs of $conv3$ layers and $conv4$ layers to build the detection feature map with both high-resolution features and high-level semantic information. We first use 1x1 convolution filters to the output of $conv3$ for dimension reduction and then stack the adjacent features into different channels to generate a smaller feature map similar to YOLOv2 [21]. This down-sampled feature map is concatenated with the output of $conv4$ layers, followed by 1x1 convolution filters to match the input channels with $conv5$ layers. Our feature fusing strategy introduces little overhead and experiments show that it greatly improves the accuracy of small objects (sleeping and hand-raising).

3.4. Online Hard Example Mining

Our dataset faces large class imbalance, see Fig. 1(c). Firstly, the sleeping samples are nearly 5-24 times less than other classes, which leads to poor sleeping detection perfor-

mance using original Faster R-CNN. Secondly, hand-raising samples are quite low-resolution with various gestures and severe occlusion and more difficult to learn compared with bigger standing behaviors, which lead to class imbalances between hard and easy examples.

To address the above class imbalances problems, we utilize OHEM (Online Hard Example Mining) [13] during training. In the forward pass, we evaluate the loss of all proposals and then sort all RoIs by loss. We choose the top-K RoIs that have the highest loss and perform back-propagation based on the selected examples. The model is learned to focus on these hard examples like the uncommon sleeping examples and various hand-raising examples. Experiments show that using OHEM greatly improves the capability of models on handling difficult scenarios.

4. EXPERIMENTS

To demonstrate the effectiveness of our proposed algorithm, we conduct extensive experiments on our student behavior dataset and the selected hard-set from our dataset. For both datasets, we show the results with metrics of COCO-style [12]: mean Average Precision (mAP), AP_s, AP_m, and AP_l. The symbols AP_s, AP_m, and AP_l represent the mAP(%) of small, medium and large objects, respectively.

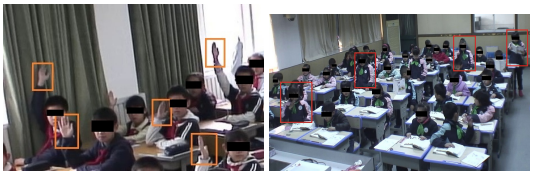
4.1. Experiments on our dataset

We perform our improved Faster R-CNN network on the student behavior dataset, including 70k hand-raising samples, 20k standing samples, and 3k sleeping samples. Our dataset comes from more than 30 different primary and middle schools in Shanghai, China. The behavior samples we captured in the dataset are quite challenging for detection due to the complex scenarios and various gestures. We use 29k images (out of 40k images total) samples for training, then validate performances on an 11k subset of the total. Original Faster R-CNN is used as the baseline for comparison to demonstrate the effectiveness of our proposed methods.

Table 1: **Experiment results of baseline and our methods on our student behavior dataset.**

	mAP(%)	AP_s	AP_m	AP_l
baseline	54.2	1.8	39.1	56.8
+feature fusion	55.3	9.1	42.1	57.9
+scale-aware head	56.5	3.8	43.4	59.1
+OHEM	56.2	11.0	42.0	58.7
ours	**57.6**	**22.4**	**45.3**	**59.6**

As shown in Table 1, our model achieves better performances than original Faster R-CNN. These ablation results show continuous improvements of our method. The final mAP has been increased by 3.4% combining all these three methods compared with baseline. Moreover, AP_s has been increased greatly by 20.6%, which reflects the effectiveness of our methods.



(a) **Detection results of baseline**

(b) **Detection results of ours**

Fig. 4: **Examples of detection results obtained on some images from the test-set.** The top and bottom rows show the detection results of the baseline and our methods, respectively. Compared with the results of baseline, our methods can detect more behaviors with low-resolution and severe occlusion.

4.2. Experiments on selected hard-set

We construct a selected hard-set from our student behavior dataset to demonstrate the usefulness of our methods for hard scenes. Specifically, we select 4.7k samples out of 11k test-set according to the relative object size (only occupy 0.1% below part of the whole image), which mainly contain sleeping samples and low-resolution hand-raising samples. As shown in Table 2, our proposed methods improve the mAP by 2.2% and are especially effective in solving this extremely hard scene.

Table 2: **Experiment results of baseline and our methods on the selected hard-set from our dataset.**

	baseline	ablations			ours
feature fusion?	–	✓			✓
scale-aware head?	–		✓		✓
OHEM?	–			✓	✓
mAP(%)	25.3	26.2	26.5	26.9	**27.5**

5. CONCLUSION

We present an improved Faster R-CNN network for student behavior detection in real classroom scenes to improve teaching quality. A scale-aware detection head using different dilation rates is proposed to detect objects of various sizes. A new feature fusion strategy is introduced for detecting the rather low-resolution objects. OHEM is used during training to alleviate the class imbalance in our dataset. The integration of these improvements achieves impressive results in real classrooms.

6. REFERENCES

[1] Maria Ofelia Pedro, Ryan Baker, Alex Bowers, and Neil Heffernan, "Predicting college enrollment from student interaction with an intelligent tutoring system in middle school," in *Educational Data Mining*, 2013.

[2] Maria Ofelia San Pedro, Jaclyn Ocumpaugh, Ryan S Baker, and Neil T Heffernan, "Predicting stem and non-stem college major enrollment from middle school interaction with mathematics educational software.," in *EDM*, 2014, pp. 276–279.

[3] Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau, "Fall detection from human shape and motion history using video surveillance," in *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*. IEEE, 2007, vol. 2, pp. 875–880.

[4] Ling Shao, Ling Ji, Yan Liu, and Jianguo Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 438–445, 2012.

[5] Janez Zaletelj and Andrej Košir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP journal on image and video processing*, , no. 1, pp. 80, 2017.

[6] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.

[7] Jiaojiao Lin, Fei Jiang, and Ruimin Shen, "Hand-raising gesture detection in real classroom," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6453–6457.

[8] Wen Li, Fei Jiang, and Ruimin Shen, "Sleep gesture detection in classroom monitor system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7640–7644.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[13] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[16] Bharat Singh and Larry S Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[17] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang, "Scale-aware trident networks for object detection," *arXiv preprint arXiv:1901.01892*, 2019.

[18] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[19] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.