SLEEP GESTURE DETECTION IN CLASSROOM MONITOR SYSTEM

Wen Li, Fei Jiang, and Ruimin Shen

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

ABSTRACT

This paper proposes a novel method to detect sleep persons in real classroom scenes, which is useful for detecting the attention of students. There are several challenges for sleep gesture detection, including occlusion, various sleep gestures, interference terms with similar features like writing, and small sleep gesture targets. To solve these challenges, we first build a sleep gesture dataset from hundreds of real classes among schools in Shanghai. Second, to detect sleep gestures we propose a modified R-FCN integrated with feature pyramid and deformable convolution. Moreover, we design an efficient local multiscale testing algorithm to address small sleep gesture detection. Experiments on our sleep dataset have shown that our approach significantly outperforms the basic R-FCN and reaches 0.74 AP@0.5. Especially, for small-size sleep gestures, our method gets an impressive improvement at 90% on our hard-samples dataset with just little additional time consumption. Besides, experiments on a large number of real video streams have proven our algorithm is applicable in real classroom scenes.

Index term– sleep gesture detection, R-FCN, feature pyramid, deformable convolution, local multiscale testing

1. INTRODUCTION

Human behavior analysis is quite useful, such as the driver state analysis during driving [1], the pedestrian behavior analysis at intersection [2] and hand gesture recognition for human-computer interaction [3]. Recently, vision-based deep learning methods are used to analyze student behavior in classroom [4, 5], which faithfully reflect the teaching atmosphere and quality. In this paper, we focus on sleep gesture detection, shown in **Fig. 1**. Several challenges exist, such as similar gestures, various sleep gestures and small sleep gestures.

Based on the researches on other gestures detection [4, 6], two kinds of algorithms are usually utilized: pose estimation algorithms and object detection algorithms. For pose estimation algorithms, [7] first employs Deep Neural Networks (DNNs). [8] proposed an innovative and robust person key points detection algorithm and [9] proposed an algorithm with a different methodology but also get good results. For object detection algorithms, there are two-stage methods like R-FCN [10] and one-stage methods like SSD [11].

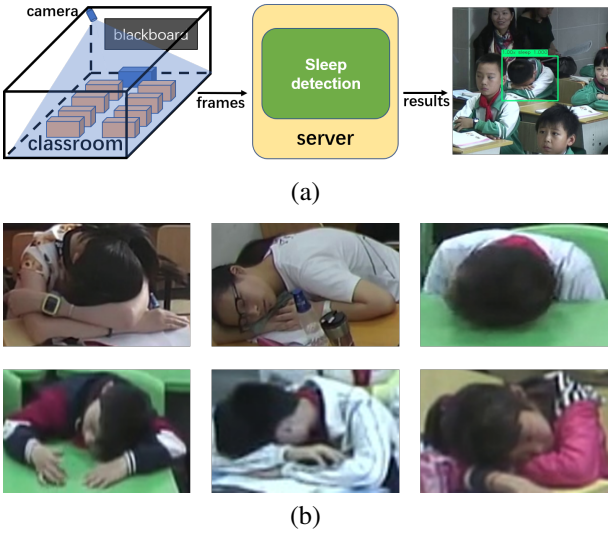

Fig. 1. **Sleep gesture detection.** (a) shows our approach uses video streams of live cameras in class as input to the local server or the cloud server. Then our algorithm will output the results for further analysis. Our approach surpport real-time analysis. (b) shows various sleep gestures in our dataset.

For pose estimation, the algorithm will generate human key points like heads, shoulders, and knees. However, in real classroom scens, the bodies of students are often occluded with many key points missing. Besides, due to similar behaviors such as writing or just head-down, we hardly find a proper mapping for these key points to specific gestures.

For object detection, two-stage methods with higher accuracy are more suitable for our sleep detection task. Two-stage object detection algorithms such as Faster R-CNN [12], R-FCN [10], and Mask-RCNN [13] are recently proposed, which achieve impressive detection results on public datasets like PASCAL VOC [14]. By testing on several existing two-stage algorithms, we choose R-FCN [10] as our basic back-

The work was supported by NSFC (No. 61671290), the Key Program for International S&T Cooperation Project of China (No. 2016YFE0129500), Shanghai Committee of Science and Technology (No. 17511101903), and China Postdoctoral Science Foundation (No. 2018M642019).

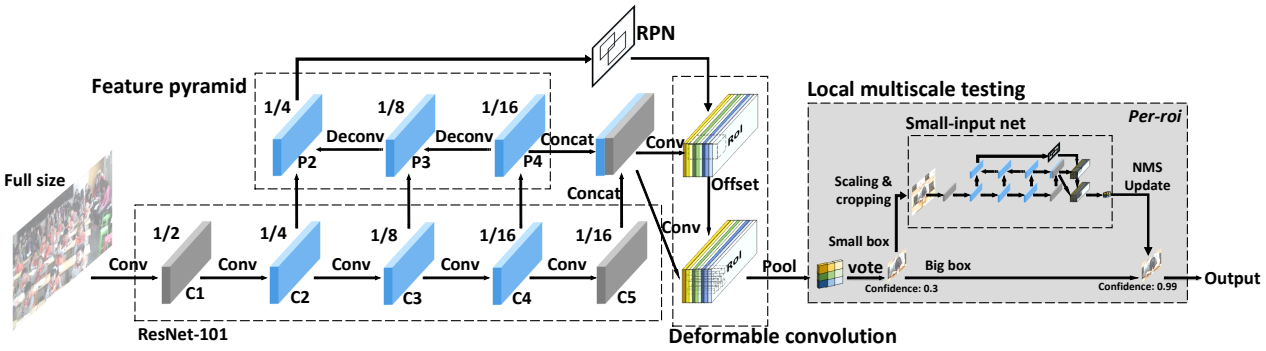

Fig. 2. **Overall structure of our approach.** A feature pyramid is built on conv2, conv3 and conv4, which is used for RPN's input. Deformable convolution is used to do deformable position sensitive ROI pooling. Local multiscale testing addresses discrimination of small targets.

bone.

Based on the above analysis, we propose a novel object-based algorithm to efficiently detect the sleep ones in real classes. First, we label the sleep gestures with bounding boxes from real class videos, and built a sleep gesture dataset. Second, we modify the R-FCN [10] detection network by employing a feature pyramid to exploit detailed features from former layers to better distinguish sleep gesture from other gestures, and using deformable convolution networks [15] method in the PSRoiPooling part to make the detection network more adaptable to various sleep gestures, then train the network on our dataset. Third, we use a local multiscale testing (LMT) algorithm to address small sleep gesture detection problem. The exprimental results on our real dataset have demonstrated the effectiveness and effciency of the proposed algorithm.

2. OUR APPROACH

In this section, we first introduce the overall architecture of the proposed algorithm for sleep gesture detection, and then explain the improvement parts utilized in the network structure. The proposed local multiscale testing strategy is presented thereafter.

2.1. Overall Architecture

The overall architecture is shown in **Fig. 2**, which can be divided into two parts, i.e., the network for sleep gesture detection and strategy of local multiscale testing. The detection network is based on R-FCN [10] object detection neural network, and ResNet-101 [16] is used to extract features. ResNet-101 [16] has 5 residual blocks, and each block is marked as {C1, C2, C3, C4, C5}. We modify the original R-FCN [10] to make our detection more precise: in region proposal network(RPN), instead of using the C4 as input, we build a feature pyramid which combines C4, C3, C2

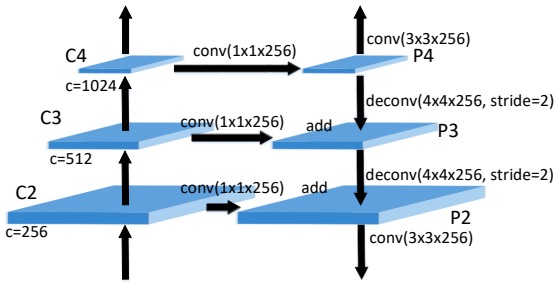

Fig. 3. **Feature pyramid structure.** Left part is 3 blocks of the ResNet [16] backbone, and right part is the feature pyramid. Here we use deconvolution instead of nearest interpolation to upsample the feature map.

as the RPN's input, and P4 is also concatenated with C5 to enrich the feature extraction; and the original position sensitive Roi pooling is integrated with deformable convolution network (DCN). The second part is a novel local multiscale testing algorithm to address the small sleep gesture detection problem, which is both time and memory efficient.

2.2. Feature Pyramid

Feature pyramid has many approaches, for example, discretized image pyramid [17] and multiple feature maps [18]. Combined with multiscale detection [19], [20] proposed a feature pyramid network (FPN) for object detection and get significant performance improvement.

We want to exploit features extracted at previous layers to have more details and we build a feature pyramid shown in **Fig. 3**. Especially, different from FPN [20], we use deconvolution [20] instead of nearest neighbour or bilinear interpolation to upsample, because we think naive upsampling

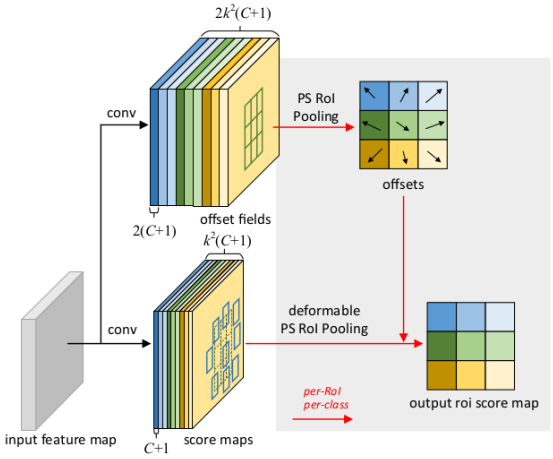

Fig. 4. **Deformable position sensitive RoI pooling [15].**

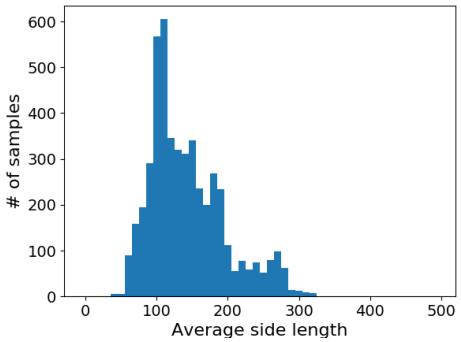

Fig. 5. **Distribution of average side length.** Average side length is the mean value of the height and width of a roi. Most samples' average side length are around 150 pixels.

ways may not correctly express the spatial relation between different levels of features. A major challenge in sleep gesture detection is that many minor differences could affect the ground truth, for example, some facial expressions and opening eyes. However, features extracted at deeper layers often represent high level semantic meaning, which may omit those minor differences. So the features from previous layer is important in our sleep gesture detection.

2.3. Deformable Convolution

Deformable convolution addresses spatial transformation problems of object features. A problem we are facing is sleep gesture could vary a lot, which can be decomposed into spatial transformations. We also notice that the bounding boxes of the detected targets are sometimes inaccurate, that is though the detection is correct, but the position and the size of the bounding box is confusing. So we apply deformable position sensitive RoI pooling [15] in R-FCN [10], shown in **Fig. 4**.

2.4. Local Multiscale Testing (LMT)

In training samples, different size of bounding boxes is not distributed equally, as shown in **Fig. 5**. The unequal distribution causes the network not sensitive to those targets of small sizes. In practice, usually a hard confidence thresh is set to distinguish good or bad detected targets. But this method will omit large percent of small true positives whose size is small than most training samples. Multiscale algorithm works well for small targets, for example in tiny face detection [21] and in pedestrian detection [22]. But the naive multiscale algorithm is both time and memory intensive.

In our experiment we find most small true positives have low confidences, so we propose a Local Multiscale Testing algorithm. Our algorithm process output boxes of different sizes separately. We set a $size$ thresh $\boldsymbol{S}$. For targets whose $size$ is larger than $\boldsymbol{S}$ we process them with a hard thresh $\mathbf{thresh_h}$, only targets of confidence bigger than it will be kept. Here, $size = (height + width)/2$.

For those targets whose $size$ is smaller than $\boldsymbol{S}$, we have a soft thresh $\mathbf{thresh_s}$ and a hard thresh $\mathbf{thresh_h}$. First, we keep those targets whose confidence is bigger than $\mathbf{thresh_s}$. And for these kept targets, we set a same sleep detection network, but with a small input size (1/4, for example). For each kept target roi_i, we rescale the image to enlarge the $size_i$ of the roi_i to the average value $\overline{size}$ in traing set:

$$image_i = \mathbf{rescale}(image, ratio = \overline{size}/size_i) \quad (1)$$

Then according to the input size $size'$ of the small-net we set, we crop $image_i$ to get the image patch with the center of roi_i at the center of the image patch:

$$patch_i = \mathbf{crop}(image_i, center(roi_i), size') \quad (2)$$

Then we collect these patches and feed them to the small-input net to get the new results. We use the new results to update the original small targets by applying Non-Maximum Suppression (NMS) on the combination of them. After the update is done, $\mathbf{thresh_h}$ is applied to them, and those targets of confidence larger than $\mathbf{thresh_h}$ are kept.

This local multiscale testing algorithm will only increase detection time and memory usage slightly because the input size of the net is small, and we know that the detection time and memory ususage is $O(N^2)$, where N is the input side length. And this algorithm performs better in small sleep gesture detection, not only the recall is increased, but also the bounding box position is more accurate.

3. EXPERIMENT RESULTS

We have conducted extensive experiments on our sleep gesture dataset and real video streams to prove our approach works. Estimation on dataset uses average precision(AP) [14], and on hard samples we focus on the recall.

3.1. Our sleep gesture dataset

Our dataset contains images with bounding box annotations. The images were taken by 1080P camera in real classroom environments, and we gather our data from primary, middle and high schools in Shanghai, China, in which there are students from 7 to 18 years old. Our dataset contains 3k images which include 5k sleep gesture targets. We use a subset of 4k samples for training and 1k to test. Besides, a hard subset of test set which contains 157 images with small (less than 80 pixels) sleep gesture targets, is used to demonstrate the usefulness of LMT for small sleep gesture detection.

3.2. Sleep gesture detection in our dataset

We conduct several experiments on our dataset to prove our approach works well in real classroom condition. **Table 1** shows the detail results of our experiment. The baseline 0.608(AP@0.5) is the original R-FCN [10]. Comparision between bilinear interpolation upsample and deconvolution shows deconvolution performs better. By applying deformable convolution the result increased to 0.6713. By applying both modifications, AP@0.5 finnally reaches 0.7477. We also test Faster-RCNN with FPN, and the AP@0.5 is 0.6309. There is also a significant improvement on AP@0.7 which means more accurate bounding boxes are generated by our proposed approach. The detail result is shown in **Table 1**.

Table 1. The result of our experiments.

Algorithm	AP@0.5	AP@0.7
Faster-RCNN+FPN	0.6309	0.4705
R-FCN [10] (baseline)	0.6080	0.3904
RFCN+FP (bilinear)	0.6709	0.4432
RFCN+FP (deconv.)	0.7245	0.4636
RFCN+DCN	0.6713	0.4822
RFCN+FP+DCN (ours)	**0.7477**	**0.5506**

3.3. Local multiscale testing in the hardset

Table 2 shows AP improvements among the hard subset testing with the proposed LMT. As shown in **Table 2**, AP@0.7 is increased significantly which means the generated bounding boxes are more accurate. By setting a hard confidence thresh 0.9, 40% are detected by our modified R-FCN [10]. By applying LMT, with hard thresh 0.9 and soft thresh 0.1, 76% are detected with only additional 23% time consumption. **Fig. 6** shows a demo of LMT.

Table 2. The result of our experiments on the hard set.

Algorithm	AP@0.5	AP@0.7
R-FCN [10] without LMT	0.3280	0.2242
Ours without LMT	0.6021	0.3995
Ours with LMT	**0.6477**	**0.5223**

Fig. 6. Demo of LMT. The red bounding box marks the area which is cropped and scaled to feed the small-input net. '2.03x' in the green tag means scaling ratio is 2.03.

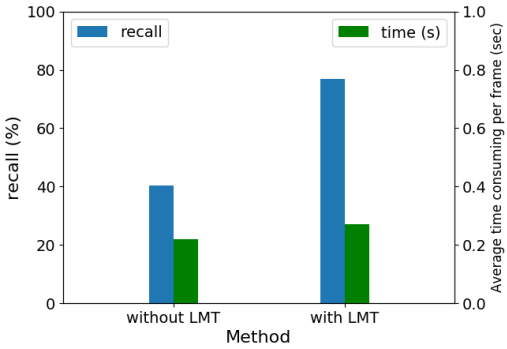

Fig. 7. Perfomance of LMT. With LMT, the recall increase 90% only at a cost of 23% time consumption increase. The testing is conducted on a single nvidia GTX 1070 GPU.

The straightforward multiscale testing, which is scalling the full image, were not considered. Those small targets usually need to be scaled at around 2 to 4 times bigger, which significantly increase the memory consumptions of GPU and detection time, and it is unacceptable for real applications. **Fig. 7** illustrates the testing results by using LMT, which dramatically improve the detection performances with slightly increase of detection time.

4. CONCLUSION

We design a sleep gesture detection architecture in the real classroom environment. Feature pyramid was introduced to exploit features from both deep and shadow layers. Deformable convolution is used to make the network more adaptable to mutable sleep gestures. We also propose a local multiscale testing algorithm to address small sleep gesture targets detection. All these strategies work together to effectively detect the sleep students in real classrooms.

5. REFERENCES

[1] Sinan Kaplan, Mehmet Amac Guvensan, Ali Gokhan Yavuz, and Yasin Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[2] Mohammad Shokrolah Shirazi and Brendan Tran Morris, "Vision-based pedestrian behavior analysis at intersections," *Journal of Electronic Imaging*, vol. 25, no. 5, pp. 051203, 2016.

[3] Siddharth S Rautaray and Anupam Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[4] 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*. IEEE, 2018, pp. 6453–6457.

[5] Benchi Shao, Fei Jiang, and Ruimin Shen, "Multi-object detection based on deep learning in real classrooms," in *Pacific Rim International Conference on Artificial Intelligence*. 2018, pp. 352–359, Springer.

[6] Diogo C Luvizon, David Picard, and Hedi Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, vol. 2.

[7] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[9] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *The IEEE International Conference on Computer Vision*, 2017, vol. 2.

[10] 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.

[11] 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.

[12] 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.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *The IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," *CoRR, abs/1703.06211*, vol. 1, no. 2, pp. 3, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[18] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *The IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[19] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[20] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie, "Feature pyramid networks for object detection.," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 1, p. 3.

[21] Peiyun Hu and Deva Ramanan, "Finding tiny faces," in *The IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1522–1530.

[22] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.