HAND-RAISING GESTURE DETECTION IN REAL CLASSROOM

Jiaojiao Lin, Fei Jiang, and Ruimin Shen

Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
johere@sjtu.edu.cn, jiang.feiabc@163.com, rmshen@sjtu.edu.cn

ABSTRACT

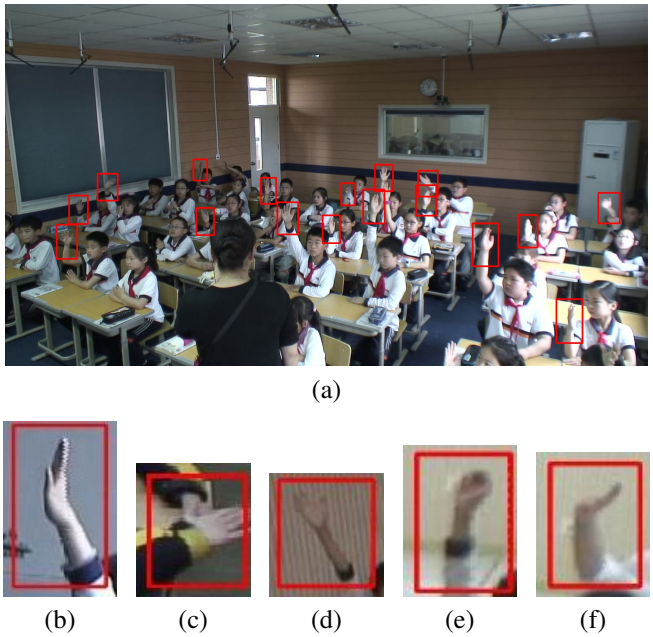
This paper proposes a novel method for hand-raising detection in the real classroom environment. Different from traditional motion detection, the hand-raising detection is quite challenging in the real classroom due to complex scenarios, various gestures, and low resolutions. To solve these challenges, we first build up a large-scale hand-raising data set from thirty primary schools and middle schools of Shanghai, China. Then we propose an improved R-FCN to solve the above-mentioned challenges. Specifically, we first design an automatic detection templates algorithm for various gestures of hand-raising detection. Second, for better detection of the small-size hands, we present a feature pyramid to simultaneously capture the detail and highly semantic features. Incorporating these two strategies into a basic R-FCN architecture, our model achieves impressive results on real classroom scenarios. After a wide test, the accuracy of the hand-raising detection achieves 85% on average, which can satisfy the real application.

***Index Terms*—** hand-raising detection, automatic templates selection, feature pyramid, R-FCN

1. INTRODUCTION

Human behavior analysis has numerous applications such as smart video surveillance [1], human-machine interaction [2], and virtual reality systems [3]. In this paper, we focus on the hand-raising detection in real classroom, aiming at improving the teaching quality.

Existing algorithms for hand-raising detection can be divided into 3 categories, i.e. trajectory based, body silhouette based and object detection based algorithms. [4] and [5] adopt a set of HMM [6] models based on frame sequences to capture the hand-raising gestures. [7] analyzes the body silhouette to detect the hand-raising motion. But these methods highly depend on the environments, and the trajectory is changed due to the movements of the camera. [8] provides a new way for the hand-raising detection, which converts such motion detection into an object detection problem. Unfortunately, [8] is based on handcrafted features, i.e. haar-like [9] features, which is not efficient in our complex scenarios, shown in Fig. 1. Motivated from [8], we also adopt an object detection algorithm, but based on automatically feature extraction techniques.



(a) (b) (c) (d) (e) (f)

Fig. 1: **Hand-raising gestures.** (a) hand-raising in a real classroom; (b)-(f) various hand-raising gestures.

Recently, Fast R-CNN [10], Faster R-CNN [11] and R-FCN [12] are three of the most representative object detection algorithms. These methods have achieved impressive results on the benchmarks [13]. However, they could not obtain optimal results in our data set. In our hand-raising data set, most of the gestures have the size smaller than 60*60 (Fig.5), and the gestures of hand-raising are various, which are quite different from the benchmarks of object detection.

In this paper, we propose a novel hand-raising detection algorithm based on the improved R-FCN to solve the challenges of the hand-raising detection in the real classroom. First, we design an automatical templates selection algorithm

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

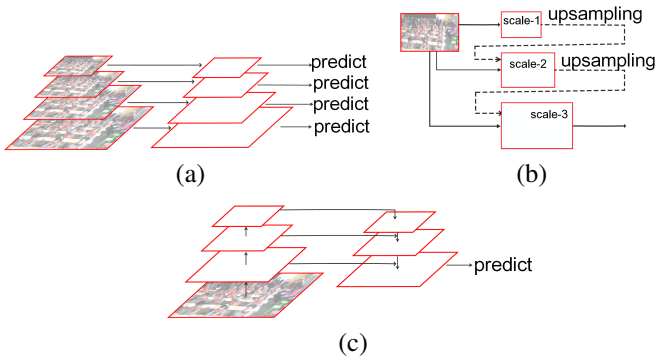


Fig. 2: **Different approaches of feature pyramid.** (a) an architecture based on a discretized image pyramid; (b) [14] extracts different scales of feature maps from different deep architecture; (c) a top-down architecture with skip connections [15]

to better detect the various gestures. Instead of choosing templates by hand in R-FCN, we run k-means++ clustering on our own training set to automatically find representative templates. Second, we utilize feature pyramid to efficiently detect the hand-raising of low resolution. We evaluate our method, including automatical templates selection, and feature pyramid learning et al. in the R-FCN system for hand-raising detection.

2. RELATED WORK

We briefly introduce the related works on region proposal algorithms and the feature pyramid architectures.

2.1. Region proposal

A region proposal method [16] [17] is one of the most important parts for object detection, since the proposals indicate all the possible locations of the object. Region Proposal Network (RPN) is the most frequently[[11]] network for proposal generation. In the last feature map of RPN, proposals are usually generated by 9 anchors at each sliding position. However, in most existing proposal algorithms, the sizes of anchors are defaulted with fixed scales and aspect ratios, which cannot adaptively match the sizes of hand-raising gestures in real classroom.

2.2. Feature pyramid architectures

To capture scale invariance, several algorithms based on feature pyramid have been proposed, as shown in Fig. 2. The first feature pyramid architecture [15], Fig. 2 (a) is based on a discretized image pyramid , which requires more computational resources but with a low speed. The second [14] and the third [15] ones are based on multiple feature maps, as shown in Fig. 2 (b)-(c). However, the second one extracts different scales of feature maps from different deep architecture, while the third one of the top-down architecture with skip connections only uses one deep architectures for all the features. Therefore, for our task, we adopt the third architecture for our feature pyramid to better detect the small sizes of hand-raising gestures.

Algorithm 1 Automatical templates selection

Input: The size of cluster, k; pairs of (w,h) in hand-raising training set, P;
Output: k kinds of templates;

```
Init k centroids in the way of k-means++;
repeat
    for all (w, h) ∈ P do
        Compute Equation(1);
        Find the nearest centroid;
    end for
    Re-compute for the new k centroids;
until Centroids not update
```

3. OUR APPROACH

We first introduce the overall architecture of the proposed algorithm. Then, we exhaustively introduce our automatic templates selection and the feature pyramid to detect the hand-raising gestures more efficiently.

3.1. Overall architecture

Our architecture is based on the R-FCN, shown in Fig. 3. We adopt the ResNet-101 [18] as the feature extract network which contains 5 blocks. We denote the outputs of these residual blocks as {C1, C2, C3, C4, C5}, respectively. Different from R-FCN, we design an automatic templates selection algorithm to detect various hand-raising gestures, as shown in RPN component of Fig. 3, and more details can be found in Section 3.2. Moreover, a feature pyramid is proposed for better detecting different sizes of hand-raising gestures. More details please see Section 3.3.

3.2. Automatic templates selection

In R-FCN, the sizes of templates are fixed. By default, 3 scales (512, 256, 128 on an input of 1000*600) and aspect ratios (1:1, 1:2, 2:1) are used for all the object detection without considering the sizes distributions of the objects. In our task, the hand-raising gestures are various (Fig. 1(b)-(f)), whose sizes are usually smaller than 60*60, see in Fig. 5. Thus, the templates in R-FCN are not appropriate to our hand-raising gestures detection.

To make the templates more suitable for our task, we automatically choose k templates by k-means++ from the bound-

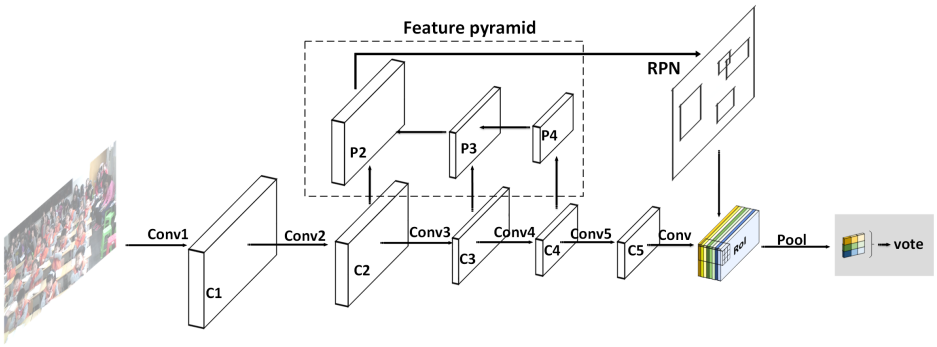


Fig. 3: **Overall architecture of our method.** A feature pyramid is built on the basis of C2-C4, where predictions will be made on the P2. The architecture includes three components, i.e., region proposal, position sensitive score maps & position-sensitive RoI pooling, and region classification.

ing boxes of the hand-raising gestures in our training set. Corresponding procedure is stated in Algorithm 1. Fig. 4 shows the average IOU with respect to k. We choose k = 9 as a good tradeoff between model complexity and high recall. Finally, 9 anchor boxes chosen as: (37,59), (44,72), (53,80), (56,96), (67,105), (75,128), (91,150), (115,184), (177,283) on an image with a size of 1000*600. The selected templates are significantly different from hand-picked anchor boxes as the R-FCN[12] used.

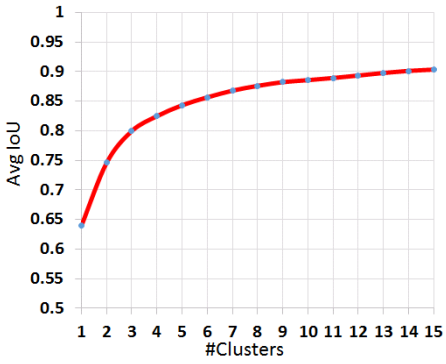


Fig. 4: **Changes of average IOU value w.r.t. the number of clusters.** The number of clusters k=9 is a good tradeoff between model complexity and high recall.

We define a new distance metric instead of the standard Euclidean distance:

$$d(box, centroid) = 1 - IOU(box, centroid) \qquad (1)$$

Here, IOU(Intersection over Union) denotes the overlap rate of anchors and the ground-truth bounding boxes in the training set.

3.3. Feature pyramid

As we can learn from Fig. 5 that the low resolution gestures are the majority of hand-raising gestures in real classroom. Thus, our detection on hand-raising need to take advantage of multi-scale in consideration of low-resolution gestures. By using the feature pyramid, our method performs better on the hand-raising data set compared to the origin R-FCN. We implement the feature pyramid by the combination

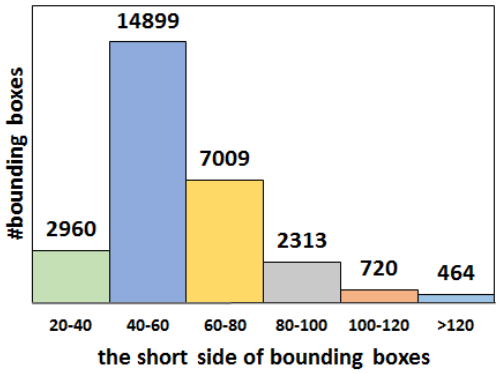


Fig. 5: **Distribution of bounding boxes for hand-raising gestures.** Most of bounding boxes are between 40 and 80 pixels.

of the bottom-up pathway and the top-down pathway. The bottom-up pathway (Fig.2(c) left) is the feed-forward computation of the ResNet-101, which computes a feature hierarchy consisting of feature maps at several scales. We built the feature pyramid on the layers of sharing weights, but we do not include C1 into the pyramid due to its large memory footprint, that is, we built the feature pyramid on C2-C4. The top-down pathway (Fig.2(c) right) obtained by combining different levels of feature maps. Feature maps from higher pyramid levels express more advanced semantic features, but finer features may be easily lost. Therefore, we combine advanced semantic features with detailed features by such pyramid architecture. We obtain P4, P3, P2, top-down, as [15] does. Fig. 6 shows the building block that constructs our top-down feature maps.

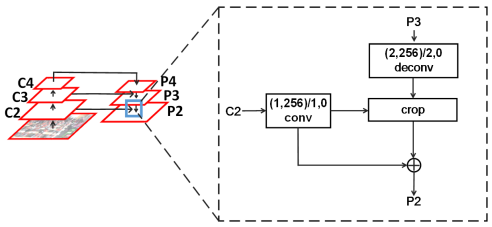


Fig. 6: **Block in top-down connection.** (1,256)/1,0 denotes a convolution layer with 256 filters of size 1*1, where the stride and padding are 1 and 0 respectively.

4. EXPERIMENT RESULT

To demonstrate effectiveness of our proposed algorithm, we conduct extensive experiments on our hand-raising data set, and then show the results with two metrics of PASCAL-style [13]: mean average precision (mAP) and Precision-Recall (P-R) curves. The mAP and P-R curves are useful measures for the performance of object detection.

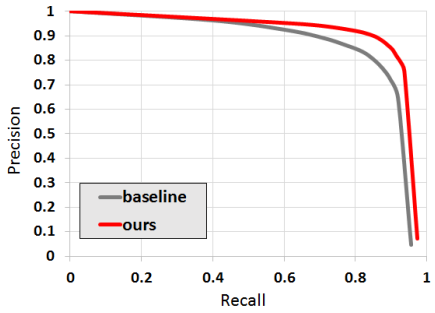


Fig. 7: **Precision-recall(P-R) curves on our data set.**

4.1. Our hand-raising data set

We perform improved R-FCN[12] network on the hand-raising data set, including 40k samples of hand-raising gestures. Our data set comes from 30 different primary and middle schools. The hand-raising gestures we captured in the data set are quite challenging for detection due to the complex scenarios, various gestures, and low resolutions. We use 28k(out of 40k total) samples for training, then report final results on a 12k subset to demonstrate effectiveness of our proposed algorithm. Original R-FCN over our data set for hand-raising detection is used as baseline for comparing results.

4.2. Hand-raising detection

For fail comparisons with original R-FCN, we run baseline and our proposed method on the same training and test set. As we see in Fig. 7, our method achieves better performance than baseline both in recall rate and precision rate.

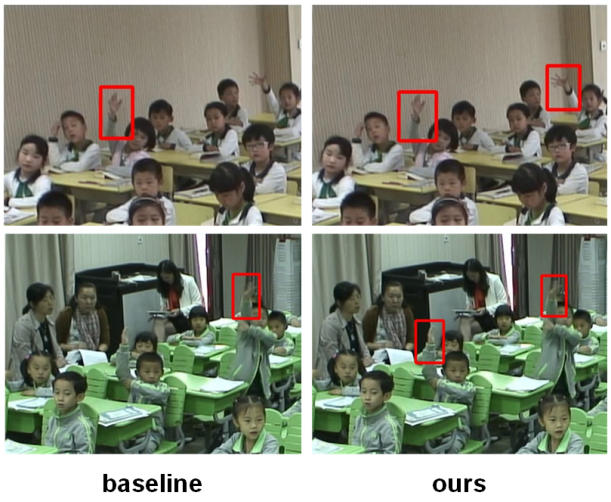


Fig. 8: **Experiment results.** The left and the right column denotes the hand-raising detection results from baseline [12] and our methods, respectively.

Table 1: **The path from baseline to our method.**

	baseline	ablations		ours
anchor cluster?		✓		✓
feature pyramid?			✓	✓
mAP(%)	83.8	86.3	87.3	**90.0**

The column named ablations of table 1 shows the mAP of the our different improvements. The design of templates clustering improves the result to 86.3% and our feature pyramid helps the mAP increased to 87.3%. The result has been increased by 6.2% using our method.

Fig. 8 shows results obtained on some of the images in the test set which compare to the results of baseline. Compared with the results from baseline, our method can detect more hand-raising with low resolution and the varied gestures which can not be detected in baseline.

5. CONCLUSION

We presented an improved R-FCN [12] network for hand-raising detection in the classroom environment, which can be utilized in the analysis of teaching atmosphere. Due to the varied hand-raising gestures, we run an adaptive algorithm on our training set to automatically pick detection templates. And we also introduce the feature pyramid in our method to simultaneously capture more detail and highly semantic features because of the majority of low resolution hand-raising gestures. The integration of these improvements achieves an impressive results in the hand-raising detection of the real classroom.

6. REFERENCES

[1] Geetanjali Vinayak Kale, Varsha Hemant Patil, and Nilanjan Dey, "A study of vision based human motion recognition and analysis," *International Journal of Ambient Computing & Intelligence*, vol. 7, no. 2, pp. 75–92, 2016.

[2] Meng Meng, Hassen Drira, Mohamed Daoudi, and Jacques Boonaert, "Human object interaction recognition using rate-invariant shape analysis of inter joint distances trajectories," in *Computer Vision and Pattern Recognition Workshops*, 2016, pp. 999–1004.

[3] Theophilus Teo, Mitchell Normal, Matt Adcock, and Bruce H. Thomas, "Data fragment: Virtual reality for viewing and querying large image sets," in *Virtual Reality*, 2017, pp. 327–328.

[4] Monowar Hossain and Michael Jenkin, "Recognizing hand-raising gestures using hmm," in *Computer and Robot Vision, 2005. Proceedings. the Canadian Conference on*, 2005, pp. 405–412.

[5] Bill Kapralos, Andrew Hogue, and Hamed Sabri, "Recognition of hand raising gestures for a remote learning application," in *Eight International Workshop on Image Analysis for Multimedia Interactive Services*, 2007, p. 38.

[6] Thad Starner and Alex Pentl, "Visual recognition of american sign language using hidden markov models," *International Workshop on Automatic Face & Gesture Recognition*, pp. 189–194, 1995.

[7] Xiaodong Duan and Hong Liu, "Detection of hand-raising gestures based on body silhouette analysis," in *IEEE International Conference on Robotics and Biomimetics*, 2009, pp. 1756–1761.

[8] Hong Liu and Dengke Gao, "Haar-feature based gesture detection of hand-raising for mobile robot in hri environments," .

[9] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *International Conference on Image Processing. 2002. Proceedings*, 2002, pp. I–900–I–903 vol.1.

[10] Ross Girshick, "Fast r-cnn," *Computer Science*, 2015.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in Neural Information Processing Systems*, 2016.

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

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

[15] C Farabet, C Couprie, L Najman, and Y Lecun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[16] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2014.

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