HAND-RAISING GESTURE DETECTION IN REAL CLASSROOM

Jiaojiao Lin, Fei Jiang, and Ruimin Shen
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

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 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.

CONTACT

Jiaojiao Lin
Email: johere@sjtu.edu.cn

MOTIVATION

We presented an improved R-FCN network for hand-raising detection in the classroom environment (Fig. 1), which can be utilized in the analysis of teaching atmosphere.

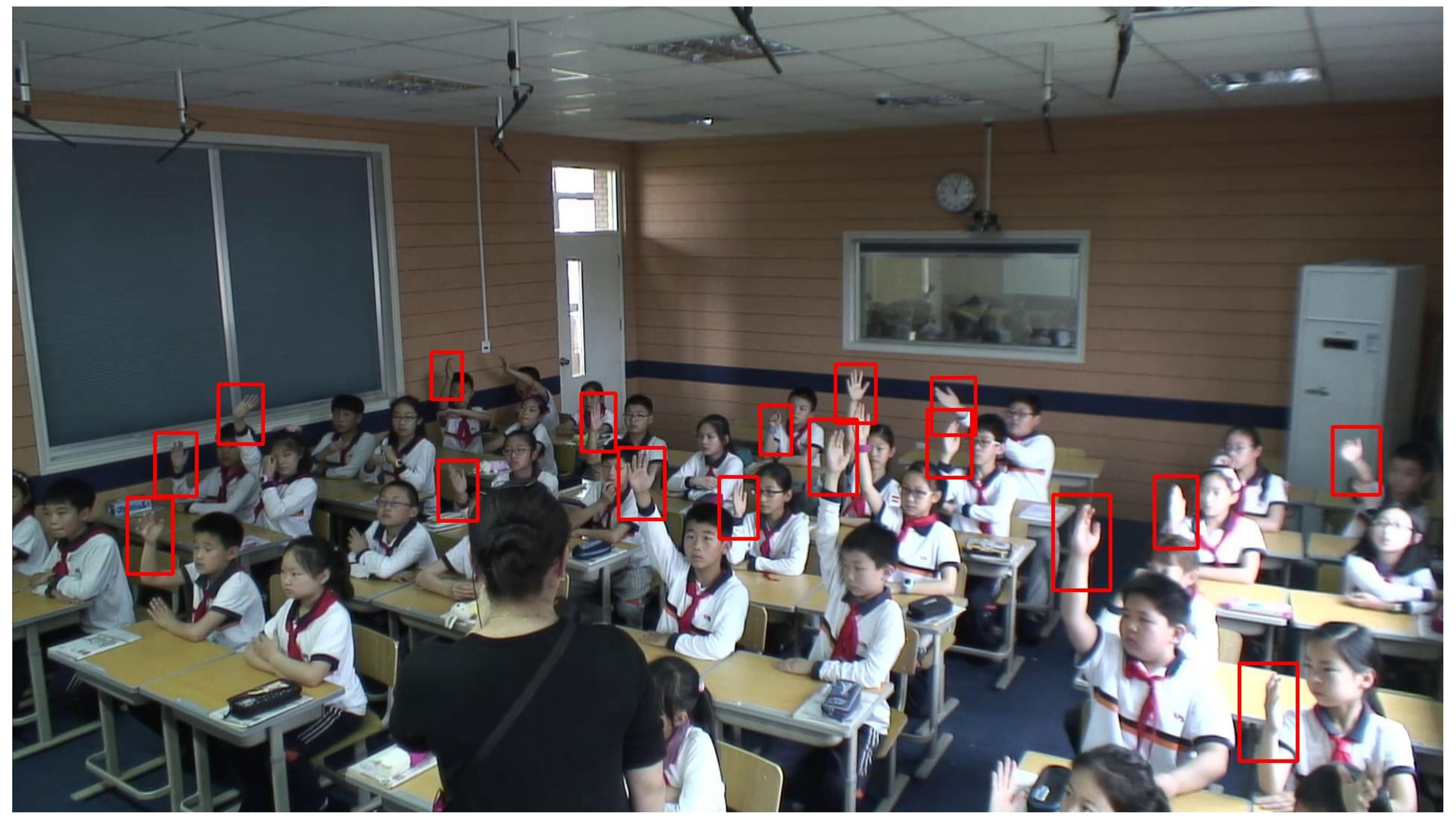

Fig.1 Hand-raising in a real classroom

OUR WORK

Different from traditional motion detection, the hand-raising detection is quite challenging in the real classroom due to:

a) Complex scenarios (Fig. 1)
b) Various gestures (Fig. 2)
c) Low resolutions (Fig. 3)

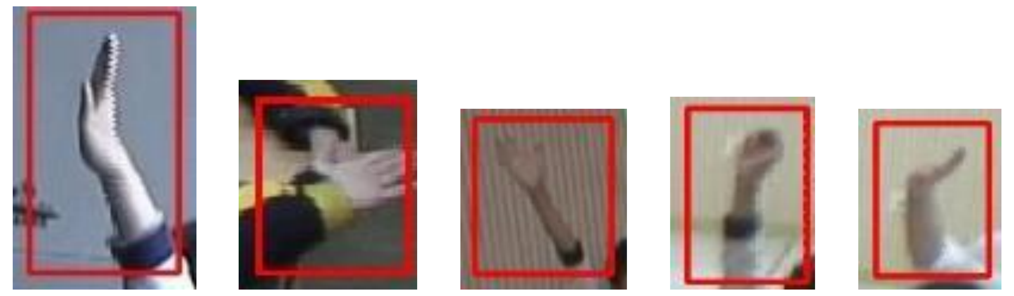

Fig.2 Various hand-raising gestures

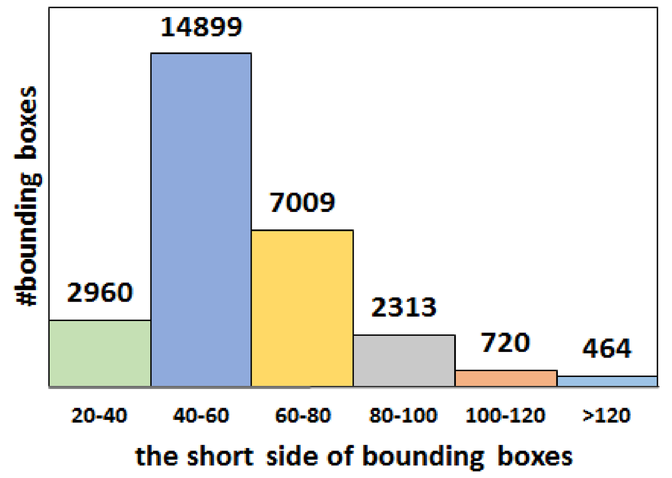

Fig.3 Distribution of bounding boxes for hand-raising gestures

Therefore, we elaborately design a network for hand-raising detection (Fig. 4):

a) We build up a large-scale hand-raising dataset from thirty primary schools and middle schools, including 60k samples of hand-raising gestures.
b) We automatically choose k templates by k-means++ from the bounding boxes of the hand-raising gestures in out training-set (Algorithm 1).
c) We build the feature pyramid on the layers of sharing weights to simult-aneously capture more detail and highly semantic features (Fig. 5).

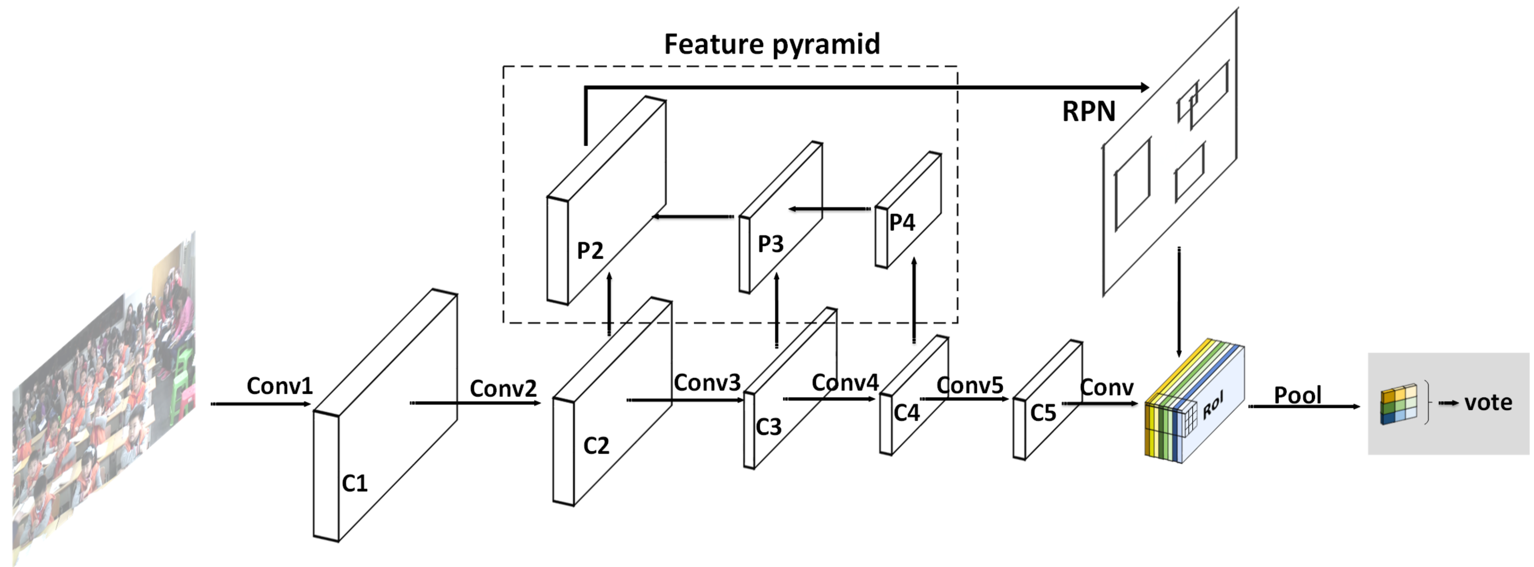

Fig.4 Overall architecture of our method.

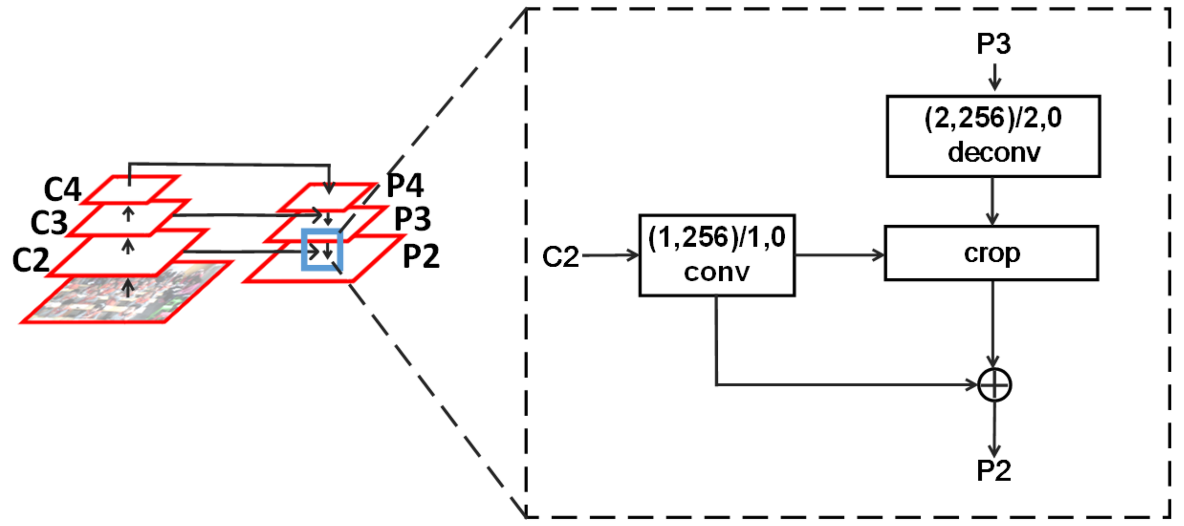

Fig.5 Block in top-down connection.

Algorithm 1 Automatically 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:
    $d(box, centroid) = 1 - IOU(box, centroid)$
    Find the nearest centroid;
  end for
  Re-compute for the new k centroids;
until Centroids not update

Experiment

For fair comparisons with original R-FCN, we run baseline and our proposed method on the same training and test set. As shown in Fig. 6, our method achieves better performance than baseline both in recall rate and precision rate.

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

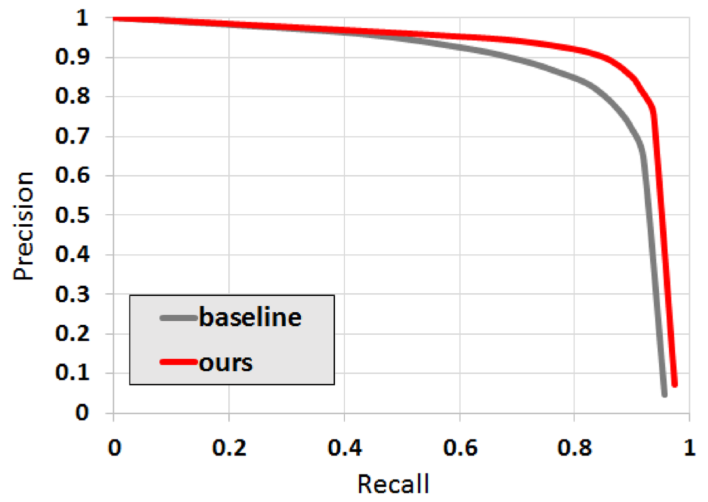

Fig.6 P-R curves on our data set

baseline **ours**