Proceedings of Machine Learning Research 95:470-485, 2018 ACML 2018

Who Are Raising Their Hands? Hand-Raiser Seeking Based on Object Detection and Pose Estimation

Huayi Zhou SJTU_ZHY@SJTU.EDU.CN
Fei Jiang JIANGF@SJTU.EDU.CN
Ruimin Shen RMSHEN@SJTU.EDU.CN
Department of Computer Science and Engineering, Shanghai Jiao Tong University

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

In this paper, we propose an automatic hand-raiser recognition algorithm to show who raise their hands in real classroom scenarios, which is of great importance for further analyzing the learning states of individuals. To recognize the hand-raisers, we divide the hand-raiser recognition into three subproblems, including hand-raising detection, pose estimation, and matching the raised hands to students. Several challenges exist while dealing with the above-mentioned subproblems, such as low resolution of the back row for keypoints detection, the motion distortion caused by hand raising in pose estimation, and various complex situations for matching. To solve these challenges, we first adopt an improved R-FCN algorithm for hand-raising detection, whose effectiveness has been demonstrated. Secondly, we present a novel PAF-based pose estimation algorithm for detecting keypoints of human bodies. The proposed PAF adds scale search and modified weight metric to adapt to the real and complex scenarios. Specifically, scale search improves the detection effect at low resolution by pooling human characteristics in different sizes of pictures, and modified weight metric reasonably utilizes the directional vectors of possible limb connections to optimize the case of motion distortion. Thirdly, a heuristic matching strategy based on the location of hand-raising and keypoints information is proposed to recognize the hand-raisers. Experimental results on six teaching videos in real classrooms have demonstrated the efficiency of the proposed algorithm, and 83% recognition accuracy indicates the potential applications in real classrooms.

Keywords: raising hand detection, keypoints detection, matching strategy

1. Introduction

Recently, with the rapid development of artificial intelligence and deep neural network, object detection and pose estimation have been flourishing. Profit from a lot of image data and computing resources, a large number of object recognition algorithms such as Fast R-CNN Girshick (2015), Faster R-CNN Ren et al. (2015), YOLO Redmon et al. (2016), SSD Liu et al. (2016) and R-FCN Dai et al. (2016) have been proposed. At the same time, a significant breakthrough has been made about the pose estimation algorithm by detecting the keypoints of the human bodies. The pose estimation is generally considered in a multi-person scene. Multi-person pose estimation mainly includes two kinds of methods. **Top-down**: First detect multiple persons, then estimate pose for each person. Representative algorithms are DeepCut Pishchulin et al. (2016), DeeperCut Insafutdinov et al. (2016), G-

RMI Papandreou et al. (2017) and RMPE Fang et al. (2017). **Bottom-up**: First detect the joints of all human bodies and then judge which person these joints belong to. The state-of-the-art methods include Newell et al. (2017) and Cao et al. (2017). With the help of these object detection and pose estimation algorithms, we can achieve the goal of confirming who are raising their hands in real classroom.

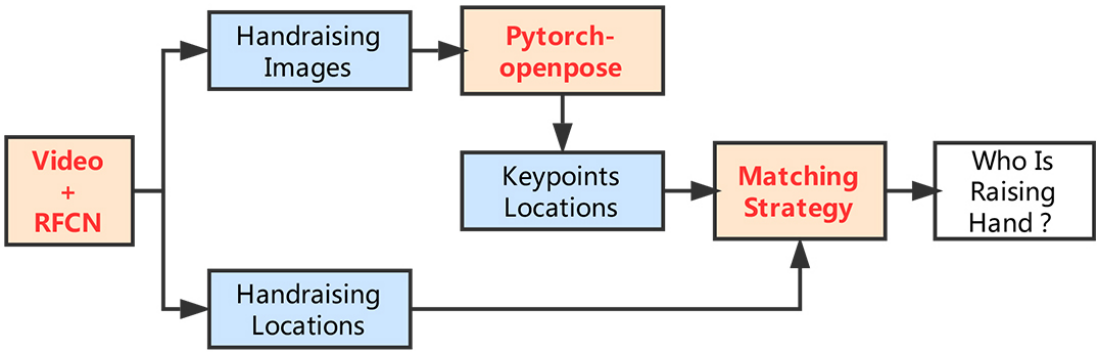

Figure 1: Flow chart of the proposed algorithm for recognizing who raise their hands.

Concretely, we choose the improved R-FCN Lin et al. (2018) to detect the raising hands in the actual teaching videos. After that, Realtime Multi-Person Pose Estimation[1] (internal PAF algorithm will be changed into our improved one) is utilized to detect keypoints of human bodies in the pictures with raising hands. Finally, according to the information of raising hand boxes and keypoints, the matching strategy can fulfil the target. The overall flow of our scheme is shown in Fig. 1. By matching the raising hands with students, we can understand their learning state of the current teacher's lecture content and the degree of cooperation with teachers. Our method is of great significance for automatic assessment of classroom atmosphere and improvement of teaching quality.

Meanwhile, many challenges in the field of computer vision will arise under the scenario of real time classroom. Specifically, our task is mainly confronted with the following two problems.

1) **Low Resolutions.** The COCO Lin et al. (2014) dataset used to train the human keypoints detection model has a higher resolution of people, about 160×260. However, in our teaching scenario, the average resolution of students is only 105×160, and the latter three rows are reduced to 80×110, which makes it difficult to detect the keypoints. Fig. 2 (a) shows some situations with low resolutions in the back row of classrooms. The left pair images show the missing keypoints detection of the raising right hand and arm, then it leads to raising hand box matching error. Paired pictures on the right miss most of the keypoints, but its a successful matching thanks to the generalized matching.

2) **Motion Distortion.**The movements of the students' raising hands may be exaggerated. A quite high lifted arm or some rare raising hands behaviors are annoying. They may change the usual position of the skeleton and bring difficulties for subsequent matching. Two instances of matching errors are displayed

[1]https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation

in Fig. 2 (b). In the left, the high lifted arm of the boy sitting in the front row is detected as the arm of a girl behand him. The right picture explains the matching errors caused by raising left hand.

(a) Low Resolution

(b) Motion Distortion

Figure 2: Examples of two main challenges in real classrooms: low rsolution and motion distortion. (a) The average width and height of the back row students are only 80×110, while the front row is 130×205. Low resolution leads to the missed or false detection of keypoints. (b) Unusual raising hand gestures can make the keypoints detection unstable and cause trouble for matching

To solve above-mentioned challenges, an improved PAF Cao et al. (2017) algorithm is proposed in this paper. We have mainly done two improvements, including the addition of scale search and a modified weight metric. The details will be introduced in section 3.

Next, a heuristic matching strategy is applied to match up the raising hands with students. The upper bound of the keypoints detection algorithm makes this assignment not easy. In many cases, the keypoints on the raising arm (including shoulder, elbow and wrist) cant be perfectly detected. That's why we need an extra raising hand detector. Fortunately, even if the keypoints are sometime incomplete, an appropriate matching strategy will achieve correct matching results. Our matching strategy takes into account almost all the situations that may be encountered in the matching process.

The rest of the paper is organized as follows. In section 2, we will explain why we choose some object recognition algorithms for detecting raising hands and some pose estimation methods to acquire keypoints. In section 3, to solve the challenges, we analyze the proposed approach in detail. And in section 4, we will present the teaching videos used for testing and show the results of contrast experiments. Finally, section 5 would be the final conclusion and some urgent problems to be solved.

2. Related Work

We will briefly introduce the algorithms involved in raising hand detection and pose estimation. The main ideas of PAF will also be emphasized.

2.1. Raising Hand Detection

Detection of raising hand gestures isnt a strange problem. Before the breakthroughs in object detection algorithm, some traditional raising hand detection algorithms such as Hossain and Jenkin (2005), Kapralos et al. (2007) and Liu et al. (2009) have emerged. However, they are easily restricted by the application environment and therefore perform bad in our rather complex real time classroom scene. Although Liu and Gao (2010) transforms raising hand detection into object detection, its accuracy is still not high enough. Until the advent of deep neural network, i.e. AlexNet Krizhevsky et al. (2012), VGGNet Simonyan and Zisserman (2014) and ResNet He et al. (2016), the deadlock of automatic object recognition was broken. An improved version of R-FCN Lin et al. (2018) based on ResNet-101 could detect raising hands in the same complex real-time classroom scenario, and it has achieved the best effects so far. Therefore, this paper will directly use Lin et al. (2018) to detect raising hands, and then match every detection box. The authenticity of raising hand boxes can only be judged manually in the subsequent statistical stage.

2.2. Pose Estimation

The development trend of pose estimation is gradually from single person to multiple persons. And multi-person pose estimation algorithms are divided into two categories: top-down and bottom-up. We will explain why bottom-up method is preferred in the real-time classroom scenario.

For single person pose estimation, detection model generated by regression method is less scalable in complicated environments. Then, new methods of extracting features by CNN have made great progress. Representative work includes Wei et al. (2016) and Newell et al. (2016). Almost at the same time, a series of multi-person pose estimation algorithms have spring up. Among them are divided into two kinds of methods: top-down and bottom-up.

Although top-down approach has achieved almost the best results on some public datasets, such as the MPII human multi-person dataset Andriluka et al. (2014) and the COCO 2016 keypoints challenge dataset Lin et al. (2014), its performance is easily affected by human detectors. Once detection of people missed, there will be no remedial measures in the follow-up. In our complex classroom scenario, it is almost impossible to detect everyone without omission. Nevertheless, the bottom-up method identifies all the human joints in the picture at first, then uses the full matching algorithm to connect them into full-body poses. Missed or false detection of keypoints may appear in results. But application based on pose estimation sometimes does not require perfect full-body poses. Especially in our scenario, incomplete human keypoints can also be utilized for matching. This is why it has achieved much better results than the top-down method. In addition, detection using bottom-up takes much less time than top-down. It is very friendly to the real time classroom.

2.3. Part Affinity Fields

The keypoints detection algorithm used in this paper is improved based on the bottom-up method PAF Cao et al. (2017) (Part Affinity Fields, a set of 2D vector fields). More details can be learned from the Realtime Multi-Person Pose Estimation[1] based on PAF. The original algorithm begins with a CNN architecture showing in Fig. 3.

The original image is firstly extracted by the convolution network composed of the first 10 layers of VGG-19 Simonyan and Zisserman (2014), and the feature maps F was extracted. Then, the network is divided into two major branches. the top branch predicts the confidence maps, and the bottom branch predicts the affinity fields. Each branch is an iterative prediction architecture, following Wei et al. (2016), In this way, after multiple stages iteration, the loss function converges, and we get the final Confidence Maps and PAF. Confidence Maps represent a series of possible human joints. And PAF with the directional information ensures that the whole human joints can be correctly connected. They will be used for Part Detection and Part Association, respectively. Finally, the Hungarian matching algorithm Kuhn (1955) and some optimization schemes are used to connect each body part and get full body poses of multiple people.

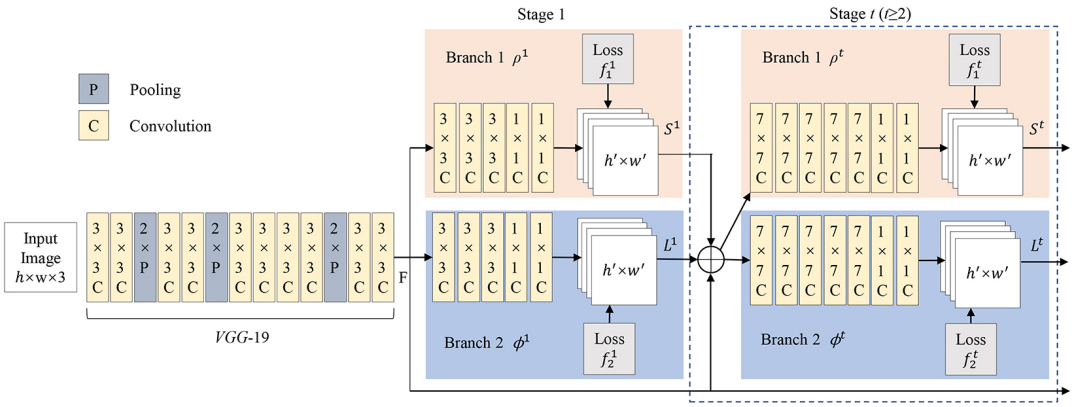

Figure 3: Architecture of a convolutional network (initialized by the first 10 layers of VGG-19 Simonyan and Zisserman (2014) and finetuned) and the two-branch multi-stage CNN. Branch 1 predicts confidence maps S^t, and branch 2 predicts PAFs L^t.

For explaining our enhanced weight metric, we will reintroduce the matching process of PAF. First of all, body part candidates are not always correct, and some false positives should be considered. When the connection of limbs is finished, unselected candidates will be abandoned, including the real body parts. Actually, body part detection candidates constitute a set $S_i = \{ d_i^j : i \in \{1, \ldots, I\}, j \in \{1, \ldots, N_i\} \}$. I is the number of real body parts. Its value is 16 in MPII Andriluka et al. (2014) and 18 in COCO Lin et al. (2014). N_i represents the number of candidates and is generated by the method of non-maximum suppression. And d_i^j indicates the coordinates of the j-th candidate of body part. The next step is to determine the possible limbs from the candidates set. We regard the candidates as nodes and the limbs as edges. This is a typical maximum weight bipartite graph matching problem Hosoya (2004).

3. Hand-raiser Recognition

This part focuses on the effective methods proposed in this paper. It mainly includes two adaptive improvements about the human keypoints detection algorithm PAF Cao et al. (2017). We add scale search to solve the problem of missed detection caused by low resolution in the back row. And a new enhanced weight metric makes the detection of rare pose perform better. It also reduces some apparent errors of the limb connection. Meanwhile, we design a heuristic matching strategy to match raising hand boxes with students accurately.

3.1. Scale Search

Figure 4: The detection effect after adding scale search at low resolution. (a) is the result without scale search. (b) uses only narrowing strategy. (c) shows the effect of adding enlarged pictures. (d) applies the complete scale search method. (e) is the comparison between the original method and our improved method.

Scale search is used to solve the problem of low resolution by pooling human features in different sizes of images. After the prediction with the CNN architecture, we will obtain some possible heatmaps of human joints and information about part affinity field. During the testing phase, we use a four scale search ($\times 0.5, \times 1, \times 1.5$ and $\times 2$) to detect the input images in various scales. The operation of enlarging pictures can help to detect the characteristics of students with smaller resolution. Narrow the picture can increase the receptive field and capture more general features. To merge the results of each scale, we first adjust the output coordinates of different scale images proportionally to the original map, and then perform averaging on the value of heatmaps and PAF outputed from network model. These results are prepared for subsequent generation of body part candidates and joint connection. Fig. 4 (a)-(d) show the effectiveness of the scale search method.

Scale search can obviously collect the effective information of small scale human bodies in the back row. After adding this trick, the keypoints in the back row are better detected comparing with the CMU-openpose[2] using original PAF algorithm. Fig. 4 (e) is the comparison results before and after adding scale search.

3.2. Enhanced Weight Metric

The enhanced weight metric is designed to better screen out possible limbs in the case of motion distortion based on part affinity fields. PAF Cao et al. (2017) scores each candidate limb using a line integral computation. Though the original method seems to make full use of all the vector fields between every two candidates, when assigning weight for each edge, it does not perform well in our scenario. There are two limitations in real classroom.

Figure 5: Many long limb connections appear before improving in the left picture. The inverse ratio of limb length is added to the weight calculation, and the problem is optimized showing in the right image.

The first limitation is that the weight metric is too unitary. In addition to considering the difference of weights caused by the orientation of the regional vectors, excessive deviations from the common sense should also be punished. For example, the length of a limb is almost

[2]https://github.com/CMU-Perceptual-Computing-Lab/openpose

the same as the width of the input picture. It has a large probability of being a wrong limb connection in this case. Without losing generality, we add the inverse ratio of limb length as a scoring penalty in weight metric. Fig. 5 shows clearly the serious mistakes of the original method in limb detection and the effect after improvement.

Figure 6: The left image is the result of the original method. No joint points have been detected of the raising hands. The right image is obtained by our modified method. The joints of the four raising hands are errorless.

The second limitation is that raising hand might be a rare gesture on the public multi-person pose estimation datasets. This makes the initial feedforward network detect fewer vector fields about lifted arms. It further reduces the weight of the real limb connection on the arm part. The left image in Fig. 6 shows the false keypoints detection of raising hand gesture. In order to reduce the impact of different quantities of vector fields, we abandon the original integral calculation and directly select a fixed number of midpoints between two candidates to get corresponding affinity fields. Then we compute the dot product of these vectors with possible limb vector and accumulate the results. Finally, this value would be added into the new weight metric. The right image in Fig. 6 gives a significantly improved detection effect.

After solving the problem of weight measurement, Hungary matching algorithm Kuhn (1955) is used to obtain the optimal matching and get a complete human skeleton. Next, we will accomplish the Multi-Person Parsing. The process of finding the optimal solution is a K-dimensional bipartite matching problem, which is NP-hard Hosoya (2004). A greedy matching method similar with the original algorithm Cao et al. (2017) will be used.

3.3. Heuristic Matching Strategy

A heuristic matching strategy is to make full use of the results of keypoints detection and seek hand-raiser for each raising hand box. After finishing raising hand detection and multi-person pose estimation, we can get some raising hand boxes and many people's full-body poses. Our goal is to find a full-body pose for each raising hand box. Then, marking the position of pose's head keypoint states that we have found who is raising hand. However,

the actual situation has encountered many problems, and the following problems are to be considered in the matching strategy.

(a) (b)

(c)

Figure 7: We can get better matching results by using a heuristic matching strategy.

(1)**There are no joints such as wrist and elbow inside raising hand box, but other joints of human body have been detected.** Under this circumstance, we cant simply give up looking for hand-raiser. we can attempt to expand the raising hand box toward the direction of the lower right or lower left, and the weight distribution of different joints should be distinctive. There are some examples in Fig. 7 (a).

(2)**A few students may raise their left hands, so the left and right arm joints should be all considered.** If having detected the students raising their right hands, we need to expand raising hand box to the low right. If left hands, the direction of box expansion is lower left. And the matching accuracy can also be improved with matching in the order of first right and then left. Successful matchings about raising left hand are shown in Fig. 7 (b).

(3)**The keypoints of human body serious overlap, and a raising hand box may be matched to several alternative full body poses.** Join some global penalties at this moment. For example, if the keypoints of the left and right arms appear in the raising hand box at the same time, it probably indicates that this is not the real hand raiser. When there are unrelated joints such as hips inside the raising hand box, it also shows that it is a wrong matching. In addition, if the position of the raising hand box is much lower than the head keypoint, it is also a sign of mismatch. Fig. 7 (c) exhibits this difficult situation.

To better understand our matching strategy, Fig. 8 shows the whole logical decision process of an example picture. In the figure, the input is the original image with coordinates

of raising hand boxes and positions of human body keypoints. And the output is an image painted raising hand boxes and matched full-body poses.

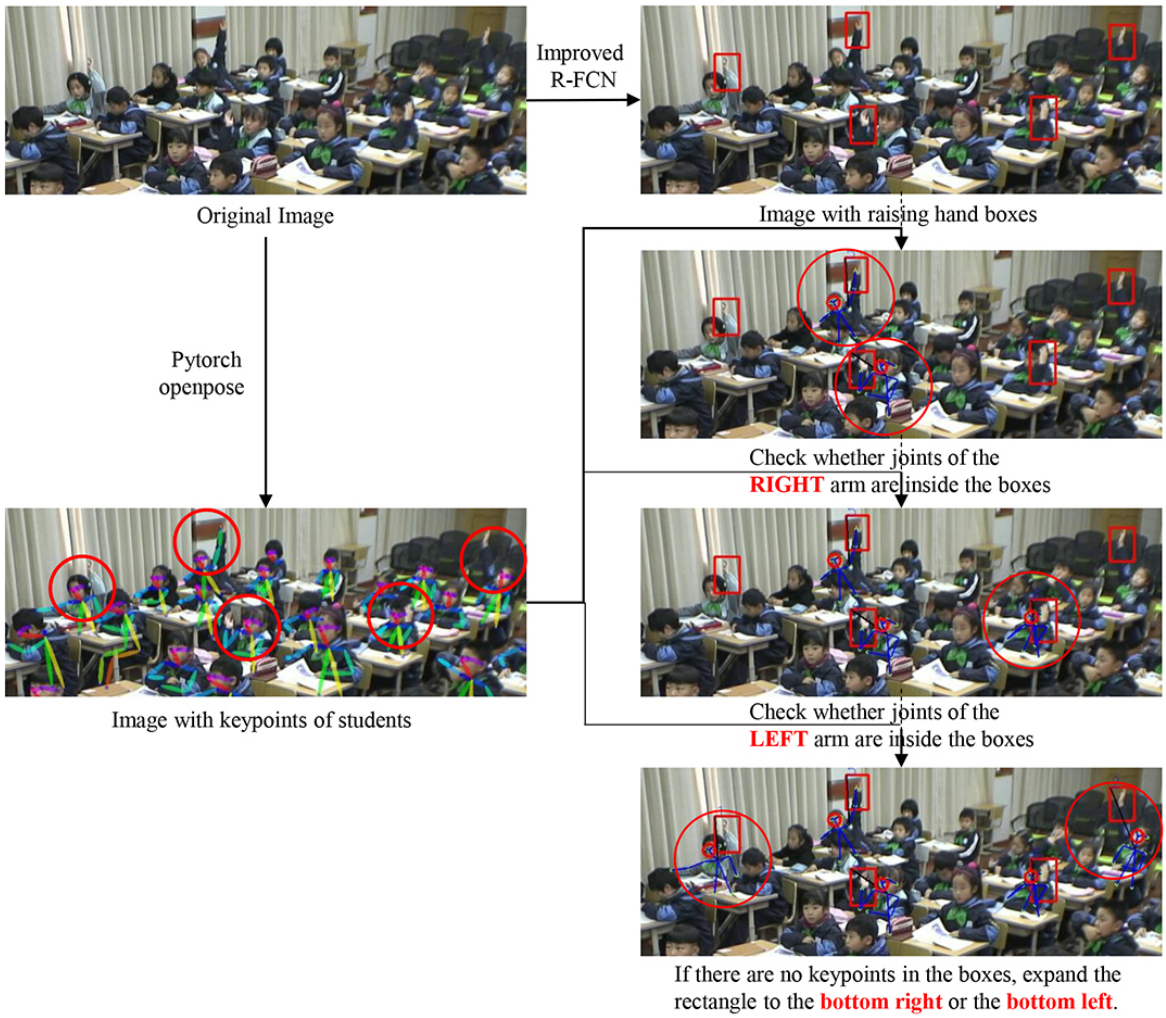

Figure 8: Thanks to the heuristic matching strategy, each raising hand box in the sample image is matched with the correct student.

In fact, pose estimation does not require high integrity of human keypoints detection in our application scenario. We sometimes tolerate errors or deficiencies in the keypoints detection of the arms. As long as a relatively complete body pose is obtained and the position of the head is determined, it is considered that the hand raiser has been discovered.

4. Experiments

In this part, we will introduce our real-time classroom videos for testing, the evaluation indicator of matching accuracy, and some contrast experiments to illustrate the effectiveness

of the proposed algorithm. The results show that our improved PAF algorithm and the proposed heuristic matching strategy have achieved good performance.

4.1. Our Test Videos

Our video dataset comes from more than 30 different primary and junior high schools in Shanghai, China. There are about 36 students in each classroom. These videos contain almost all the students in the classroom with a duration of about 40 minutes. In addition, at least three cameras are installed in each classroom. Two cameras on the left and right sides above the front wall are responsible for taking pictures of students. The camera above the back wall mainly monitors teachers.

Figure 9: The picture shows the real classroom scene. The size is 1920×1080, and about 40 students appear in the lens.

For our experiment, we selected 6 videos of different schools. Considering the influences of camera locations, we have chosen our test videos with both left and right perspectiv. Then, we cut these videos one frame every three seconds. The resolution of these frames is 1920×1080 (W $\times$ H). Fig. 9 is a frame of a real class. The pictures obtained will be checked by improved R-FCN Lin et al. (2018) whether having raising hand gestures. After testing, on average, each video will produce about 200 pictures containing raising hands. For images containing raising hands, human keypoints will also be detected using our modified PAF. These outputs will be used for subsequent matching for seeking hand-raisers.

4.2. Evaluation Indicator

A reasonable evaluation indicator is designed to judge the accuracy of matching. Specifically, suppose that we process test videos and get pictures containing raising hands. Then we detect these pictures and obtain raising hand boxes. However, the raising hand detector does not always get the real results. We have to manually count the number of real raising hand boxes. Next, the total number of matching about raising hand boxes can be calculated by program. But the number of correct matching still needs manual statistics to acquire. Finally, we get Eqn. 1. *Acc* is used to compute the matching accuracy, and *Sen* reflects the sensitivity of the matching strategy.

$$Acc = \frac{N_{correct}}{Box_{real}} , Sen = \frac{N_{total}}{Box_{all}} . \tag{1}$$

where Box_{all} is the number of detected raising hand boxes. Box_{real} represents the number of real raising hand boxes. N_{total} and $N_{correct}$ respectively indicate the total number of matched boxes and the number of correctly matched boxes.

4.3. Contrast Experiments

In order to verify the performance of the improved PAF algorithm and matching strategy, we chose the CMU-openpose[2] using the original PAF algorithm to compare with our improved PAF. In addition, we also used very simple matching strategy that ignores many special circumstances to do comparative experiments. The results are shown in Table 1.

Table 1: Test results on 6 classroom videos. L and R respectively indicate that the camera is placed on the left side and the right side above the front wall. A1 is CMU-openpose using the original PAF algorithm. A2 is our keypoints detection algorithm. M1 and M2 respectively represent simple and optimal matching strategies.

TestVideo ID-SIDE		1-L	2-L	3-L	4-L	5-R	6-R	**Total**	**Acc**	**Sen**
Box_{all}		407	201	437	953	441	228	2667		
Box_{real}		377	136	391	922	402	181	2409		
A1+M1	N_{total}	341	152	404	718	384	225	2224	–	83.39%
	$N_{correct}$	244	81	296	564	296	158	1639	68.04%	–
A1+M2	N_{total}	358	169	410	796	401	227	2361	–	88.53%
	$N_{correct}$	263	84	301	627	317	154	1746	72.48%	–
A2+M1	N_{total}	370	173	417	872	416	223	2471	–	92.65%
	$N_{correct}$	278	90	311	699	325	149	1852	76.88%	–
A2+M2	N_{total}	395	198	424	946	434	228	2625	–	**98.43%**
	$N_{correct}$	291	95	318	794	351	152	**2001**	**83.06%**	–

As shown in Table 1, our method(A2+M2) achieves the highest matching accuracy and sensitivity. Under the same video with the identical matching strategy, the original pose

estimation method(A1) obtains less total matching than the improved one(A2). It indicates that our improved algorithm can detect human keypoints better. This is surely beneficial to promoting the accuracy of matching. Thanks to our optimized matching strategy(M2), the number of correct matching has also been increased than selecting ordinary matching strategy(M1). In addition, test results of videos from both left and right side have been improved. It shows that the final accuracy will not be affected by camera locations. Furthermore, the matching accuracy of 83% demonstrates that our work is of practical significance. Some matching results are shown in Fig. 10.

5. Conclusion

To the best of our knowledge, we first propose the question about how to seek the hand-raisers in real classroom. Although the scene is very complex, mainly including two challenges: low resolution and motion distortion, we have designed our own effective solutions. We add scale search and enhanced weight metric on the basis of the original PAF algorithm to solve above-mentioned two challenges respectively. And a heuristic matching strategy is also devised to make full use of the outcomes by raising hand detection and pose estimation. Finally, the effectiveness and usability of our scheme are verified in specific examples and test videos.

In the course of the experiment, we also found some problems to be further studied. Serious occlusion is one of the troubles. If a student is almost completely blocked, 2D pose estimation is hard to detect the keypoints of him or her. Therefore, better matching results must wait for the new progress of pose estimation. With regard to matching strategy, we can not take into account all cases, so the upper bound of matching accuracy is unknown.

Acknowledgments

The work 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).

References

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

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 (NIPS)*, pages 379–387, 2016.

Figure 10: Matching results in some complex scenes,. In these four pictures, the number of correct matching / the number of raising hands is 21/26(Left), 9/12(Left), 8/10(Left), 9/9(Right) respectively. Mismatching cases are the types that are not yet perfectly resolved.

Hao Shu Fang, Shuqin Xie, Yu Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

Ross Girshick. Fast r-cnn. pages 1440–1448, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Haruo Hosoya. *Introduction to Graph Theory.* CHINA MACHINE PRESS, 2004.

M. Hossain and M. Jenkin. Recognizing hand-raising gestures using hmm. In *Computer and Robot Vision, 2005. Proceedings. the Canadian Conference on*, pages 405–412, 2005.

Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, pages 34–50, 2016.

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*, page 38, 2007.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

Harold W Kuhn. The hungarian method for the assignment problem. volume 2, pages 83–97, 1955.

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)*, pages 6453–6457, 2018.

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 (ECCV)*, pages 740–755, 2014.

Hong Liu and Dengke Gao. Haar-feature based gesture detection of hand-raising for mobile robot in hri environments. In *CAAI International Conference on Advanced Intelligence (ICAI)*, 2010.

Hong Liu, Xiaodong Duan, Yuexian Zou, and Dengke Gao. Detection of hands-raising gestures using shape and edge features. In *IEEE International Conference on Robotics and Biomimetics*, pages 1480–1483, 2009.

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 (ECCV)*, pages 21–37, 2016.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2277–2287, 2017.

George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, volume 3, page 6, 2017.

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Shih En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.