Proceedings of the 2018 IEEE
International Conference on Robotics and Biomimetics
December 12-15, 2018, Kuala Lumpur, Malaysia

Long-range Hand Gesture Recognition With Joint SSD Network

Chengming Yi*[1], Liguang Zhou*[2], Zhixiang Wang[3], Zhenglong Sun[2] and Changgeng Tan[1,†]

***Abstract*— The hand gesture recognition plays an important role in the human-computer interaction (HCI) field. Previous works mainly focused on researching this task in the shorter distance, which can only be applied in a limited scenario due to the recognition distance constrains. In this paper, we propose a method, which is named Joint Single Shot Multibox Detector (JSSD) network, to solve the hard long-distance hand gesture recognition task. The method is based on the framework of SSD. Our model employed two SSDs for object detection. One is for the head-shoulder region detection and the other is for the hand gesture bounding box proposal and classification. We validate our proposed model on real data collected from USB camera at several distance levels. The experiment results show that our method is more robust and precise in recognizing the hand gestures at a longer distance. Concretely, JSSD network is able to recognize the hand gestures captured up to 6 meters away from the camera.**

I. Introduction

Sign language is a very important communication bridge between the deaf and the societies. Generally, sign language consists of two parts, one is hand posture, represented by the position and configuration of fingers, and the other is hand gesture, which means the moving trajectory of the hand [1]. There are various works focused on hand gesture detection and recognition.

For example, a multi-class SVM [2] classifier is used to train the dataset based on the ad-hoc feature set, where the position and orientation of the fingertips is calculated. However, it's not an easy task to design a robust and reliable algorithm to adapt to the variations of environment and people [3].

The traditional algorithms require the depth information from Microsoft Kinect and the hand points and features from the Leap Motion to boost the recognition accuracy, but the hand-crafted feature descriptors are complex and require a lot of efforts to design. Besides, the above models are limited with the lighting conditions. Therefore, in order to improve the robustness and accuracy of the hand recognition algorithms, deep learning technology is used in object recognition in recent years.

*The first and second authors contributed equally to this work.

† The corresponding author of this work is Changgeng Tan.

Research supported by the Shenzhen Science and Technology Innovation Commission, fundamental research grant JCYJ20170818104502599

[1]Chengming Yi and Changgeng Tan are with Central South University, Changsha, China (email: freeape@csu.edu.cn; cgtan@csu.edu.cn). [2]Liguang Zhou, Zhenglong Sun are with the The Chinese University of Hong Kong, Shenzhen, Shenzhen, China (email: hszhoushen@gmail.com, sunzhenglong@cuhk.edu.cn). [3]Zhixiang Wang is with National Taiwan University, Taipei, Taiwan (email: wangzx1994@gmail.com).

A. Object detection based on Deep Learning

Fast-RCNN [4], Faster-RCNN [5] and R-FCN [6] are the most popular deep convolution neural networks for object detection. The detection process includes two steps: region proposal and classification. Also, there are some improved networks that do not rely on regional proposal such as SSD [7] and YOLO [8]. In general, models with region proposal have better model accuracy than models without region proposal. One of the major differences among them is that the networks that do not adopt region proposal require more hands-on parameter tuning experiences.

In this paper, the SSD [7] is used to detect the hand gestures, and the speed of SSD is faster than models with region proposal. However, the limitation of SSD is that it has a lower detection rate and recognition accuracy when hand gestures collected under the situation that the USB camera is 2m away from the hand gestures.

In addition, the problem of these neural networks is the trained model is too large for the real-time recognition. For example, the model trained by the Faster RCNN is about 200M, and the model trained by the SSD with VGG16 is 100M [9]. There are some works focused on building the light weight neural networks such as, MobileNets [10], and ShuffleNet [11], these methods can compress the model and accelerate the speed of network.

Mobilenets [10] is proposed for the mobile and embedded vision system. One of the key elements is that the depthwise separable convolution is introduced to reduce the number of computations, which is worked as the fundamental network of SSD.

Also, these methods merely discuss the hand gesture recognition from a very limited distance (2m). For this reason, in order to detect the hand gestures at a longer distance (6m), we intended to design a more convenient sign language recognition system for the deaf people to interact with the computer, and the JSSD network is proposed to prolong the recognition distance.

In summary, our main contributions of this paper are as follows:

- We proposed JSSD network, which can recognize the hand gesture up to 6 meters away from the USB camera and with a good mAP at 5 meters.
- In order to achieve high speed, the JSSD based on the Mobilenets-SSD [10] is implemented.
- We collected a dataset named `YZ Hand Gesture Dataset` based on the *Microsoft and Leap Motion dataset* [2]*, and conducted comparison experiments of

*http://lttm.dei.unipd.it/downloads/gesture/

978-1-7281-0377-8/18/$31.00 © 2018 IEEE

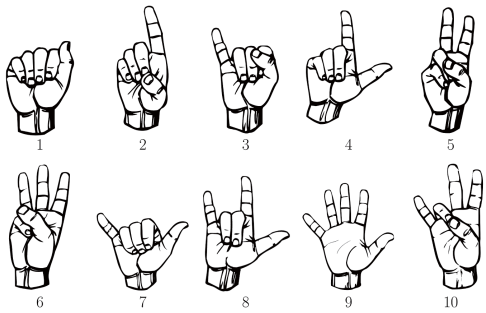

Fig. 1. Differnet kinds of hand gesture types in the American Manual Alphabet

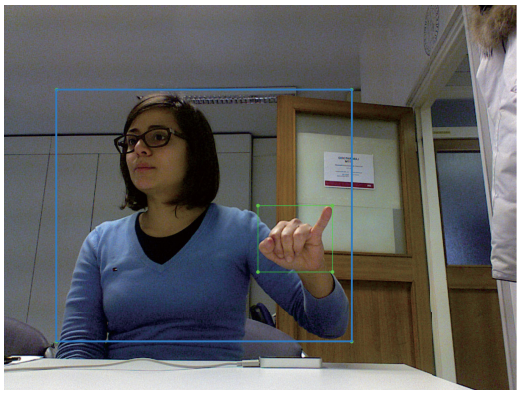

Fig. 2. The example of labeling for dataset

SSD and JSSD to verify our proposed JSSD model.

The rest of paper is organized as follows. Section II describes the SSD model for better detecting and recognizing the object and JSSD model for hand gesture recognition at a longer distance. Section III shows the experimental results and demonstrates that the JSSD can recognize the hand gestures even when the hand gestures is collected 6 meters away. Finally, the conclusions and future directions are summarized in Section IV.

II. Methodology

This section first introduces the dataset for training and testing in our system and describes the SSD model which contains the clustering of bounding boxes size for automatically selecting the optimal aspect ratio and size of rectangular. Besides, considering the limited recognition distance that SSD algorithm is able to reach, the JSSD network is developed to tackle with the tasks that require longer distance gesture recognition.

A. Dataset

There are ten types of the hand gestures in total ranging from 1 to 10 as Figure 1 shows. The hand gesture dataset collected by Microsoft Kinect and Leap Motion [2] is adopted by our paper, but we only use the RGB image of the dataset. In addition, there are 10 different types of hand gestures collected from 8 different people at different distance levels such as 1m/2m as short distance and 5m/6m as long distance. Each gesture contains 340 images at different distance levels per person, we call it `YZ Hand Gesture Dataset`. The images are captured by a common USB camera at a speed of 25fps in the lab environment without strictly illumination constrains. The size of captured images is 640×480 pixels and the type of images is RGB color. The `YZ Hand Gesture Dataset` is splitted into the training data and testing data. We randomly extract 20 images from each gesture in 1m/2m/5m dataset as the training set for a total of the $20 \times 10 \times 8 \times 3 = 4800$ images. Also, We extract 320 images from the remaining dataset at all distance levels as TEST_MX (X=1, 2, 5, 6) testing dataset. The Microsoft Kinect and Leap Motion Dataset is also splited as the training set and testing set using the leave-one-person-out approach [2], and we call the testing dataset as TEST_ONE (140 images). Therefore, there are $1400-140+4800=6060$ images in the all training set, and there are 320 images in each TEST_MX testing set. We re-label all the data using the style as shown in Figure 2. There are two bounding boxes in our training data. We first label a region of interest (ROI) including the head, shoulders, and hand gesture named head-shoulder area due to the fact that hand gestures are always near the hand-shoulder region as blue rectangular shows. After that, the hand gesture ROI is labeled as green rectangular depicts.

B. SSD—Basic Network for JSSD

a) Basic Network: SSD is a cutting-edge deep neural network for object detection. Compared with the R-CNN [12], Fast-R-CNN, YOLO network, the FPS (frames per second) of SSD is faster. Besides, the performance of the SSD and Faster-RCNN on PASCAL VOC07+12 is 74.3% and 73.2% repectively. The mAP of SSD is higher than the Faster R-CNN. The architecture of SSD model for hand gesture recognition is depicted in Figure 3, and the SSD is the basic object detection model for JSSD that we proposed. Moreover, we can also use alternative networks such as Faster RCNN, YOLO to replace the SSD network as a basic object detection model in JSSD structure .

b) Aspect ratio Clustering for SSD: The aspect ration as well as minimum and maximum size in SSD is picked by hand, which can be automatically selected by the clustering algorithms. There are two problems during training according to the YOLO 9000 [13] and DSSD [14] in detail. First, the size of the bounding box is manually picked. Although we can adjust the size more appropriately after training. However, if the better parameters of the box size are selected before training, the higher performance model can be acquired. Therefore, the k-means clustering algorithm is picked to automatically find the optimal size of bounding box. Second, because there are some different sizes of bounding boxes, thus, the Euclidean distance is not adaptable to the size variance of bounding boxes, so that the normalized distance metric in k-means is as follows:

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

with

$$\mathrm{IOU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}\,. \qquad (2)$$

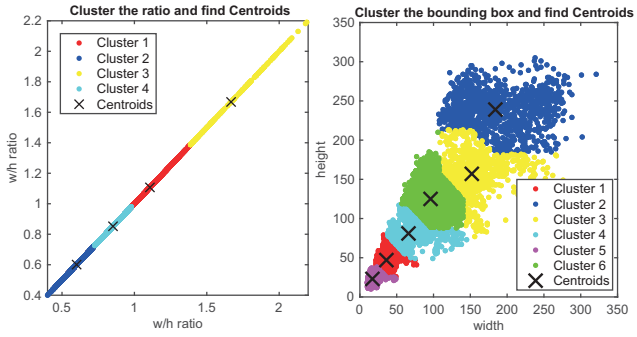

Fig. 4. Clustering results of all bounding boxes from the dataset. (left) shows the ratio of width to height are clustered into 4 categories, (right) shows the width and height of bounding box are clustered into 6 categories.

By applying above clustering rules, the aspect ration as well as the maximum and minimum size of the bounding box are clustered as Figure 4 shows.

C. JSSD network and Training

SSD is able to recognize the hand gestures collected less than 2m from the camera. However, as the distance increases, the performance of SSD becomes unreliable. In order to recognize the hand gestures at a longer distances, the JSSD network depicts in Figure 3 is proposed to increase the detection and recognition distance.

c) JSSD Architecture: The architecture of the JSSD consists of two SSDs called SSD1 and SSD2, one for extracting the head-shoulder features and the other for extracting and classifying the hand gestures. The idea is that the JSSD separate the recognition task into two subtasks when the SSD1 is for extracting the head-shoulder area and the SSD2 is for hand gesture area proposal and recognition. By combining two SSD networks, JSSD is capable of recognizing the hand gestures collected by USB camera at a longer distance like 5m, which is twice as long as a single SSD. Besides, there is a LRD Trigger Layer among the two SSD networks functioned as the Training Trigger layer. The function of trigger layers is as follows:

$$LRDTr(x) = \begin{cases} 1 & \text{if } LossTr \leq 0.05 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The equation tells us that if the head and shoulders area were not detected by the SSD1, the LRD Trigger layer will turn off so that the output of SSD1 will not be fed into SSD2. When the head-shoulder area is detected by SSD1, the LRD Trigger Layer will feed the output of SSD1 into SSD2 for hand gesture recognition.

d) Training: The loss function of SSD can be found in literature [7]. During the training process, the batch norm normalizaiton [14] technique is used for the better training as well as reaching the higher accuracy especially when the network is extremely deep. However, after the training, batchnorm becomes a time-consuming procedure in the testing phase. Therefore, in order to achieve the more efficient speed, batchnorm normalization can be removed in the model deployment phase.

e) LRD Trigger Layer Function at Inference Time: After the model is trained, we only need to set the threshold parameters of the LRD Trigger layer. For example, if the threshold is set to 0.5f, the head-shoulder detection scores greater than 0.5f can meet the requirements of gesture detection.

III. Experimental Results

We tested Fast SSD and JSSD on a laptop with the Intel I7-6700HQ CPU@2.6GH. Caffe is used to run the Joint Fast MobileNet-SSD network. The algorithm can realize the 14FPS on the laptop with above-mentioned CPU and 114FPS on the server with the single NVIDIA GPU named GTX1080Ti. When testing the trained JSSD model on the testing data, the JSSD can reach a high mAP. Besides, the algorithm can recognize the hand gesture obtained by camera up to 6 meters. Also, the system can be very robust and adapt to the different lighting conditions and backgrouds as well.

There are 320 images in TEST_M1, which means the images are captured 1m from the USB camera. Similarly, Test_M2, TEST_M5, TEST_M6 means the images captured 2m, 5m and 6m from the USB camera respectively. Each dataset have 320 images with 8 people and 10 hand gestures per person. The example images from the abovementioned dataset except the TEST_M6 is shown in Figure 5.

f) Fast-SSD with Leap Motion Dataset: As shown in Table I, The Fast-SSD based on the Microsoft Kinect and Leap Motion Dataset on TEST_ONE dataset is better than the results in paper [2], [15].

TABLE I

THE CONFUSION MATRIX RESULTS OF FAST-SSD BASED ON THE MICROSOFT KINECT AND LEAP MOTION DATASET ON TEST_ONE DATASET. BOLD ELEMENTS REPRESENT TRUE POSITIVE, WHILE OTHER CELLS SHOW FALSE POSITIVE WITH FAILURE RATE GREATER THAN 1%.

Type	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G2	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G3	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G4	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00
G5	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
G6	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
G7	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00
G8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
G9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.98**	0.00
G10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**

A. Fast-SSD

In the first experiment, the four sets of testing dataset are used to test the performance of the Fast-SSD300 model. The results show that the Fast-SSD300 works well on the TEST_ONE dataset whose mAP is 0.936. However, the Fast-SSD300 has the limited ability to distinguish and recognize the hand gestures in TEST_M1, TEST_M2, and TEST_M5, and the mAP for these three testing sets are 0.436, 0.111, and 0 respectly. From the Table II, we can find that the Fast-SSD300 is almost perfect in TEST_ONE dataset with

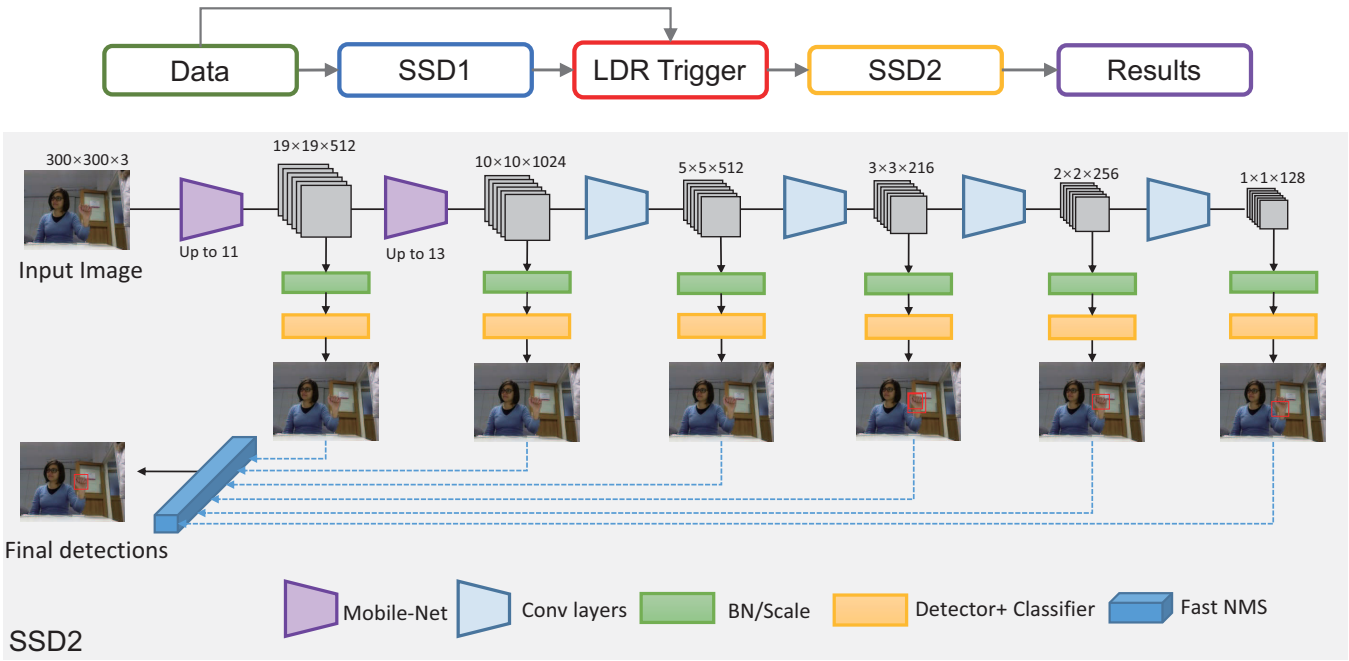

Fig. 3. Network structure of proposed JSSD model, the upper side is JSSD network pipeline, the bottom side within gray area is the network structure of SSD2 network, which is similar to SSD1 network except that SSD1 model is for head-shoulder region proposal while SSD2 model is for hand gesture detection and recognition.

Fig. 5. The example images from dataset TEST_MX (X = 1, 2, 5)

mAP 0.936, which means when the input hand gestures are collected in 1m from the camera, the system can easily recognize the input hand gestures. However, the mAP of Fast-SSD300 on the TEST_M2 is dramatically decreased to 0.111. It suggests that as the collecting distance of the dataset increases, the system performance dramatically decreases into a poor level. Therefore, the JSSD is introduced to realize recognizing the hand gestures collected at a longer distance.

B. JSSD

In the rest four experiments, the JSSD is proposed to tackle with the distance issue and can recognize the hand gestures at a longer distance. The JSSD is trained with four different input image sizes. For those input image sizes with the 224× 224, 256×256, and 300×300, there's no obvious difference between the mAP. The performance of JSSD300, JSSD256, JSSD224 at 5m are surprisingly well with the mAP 0.868, 0.864, and 0.822. However, the performance drops quickly when the distance increases to 6m that the mAP reaches 0.593, 0.585, 0.572 respectively. However, we can found that when the input image is the largest one (512×512), the model is not able to recognize the hand gestures at 5m or 6m from the camera.

C. Error Analysis

The experimental results show that the proposed JSSD can recognize the hand gestures collected 6m from the camera. However, the weakness of the proposed model is that the model does not have a good performance in the TEST_M1 dataset. Only the JSSD512 reaches the optimal performance

TABLE II

THE EXPERIMENTAL RESULTS, FAST-SSD 300 MEANS THE INPUT IMAGE SIZE IS 300×300. SIMILARLY, THE JSSD-512, JSSD-512,JSSD-300, JSSD-256, JSSD-224 MEANS THE INPUT IMAGE SIZE IS 512×512, 300×300, 256×256, AND 224×224

Method	Dataset	mAP	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Fast-SSD300	TEST_ONE	0.936	0.97	0.88	0.98	1.00	0.82	0.77	1.00	0.98	1.00	0.96
	TEST_M1	0.436	0.40	0.61	0.08	0.71	0.38	0.09	0.62	0.61	0.54	0.32
	TEST_M2	0.111	0.00	0.00	0.14	0.23	0.04	0.14	0.07	0.23	0.08	0.18
	TEST_M5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
JSSD512	TEST_ONE	0.924	0.97	0.86	0.93	0.93	0.91	0.95	0.94	0.94	0.83	0.98
	TEST_M1	0.814	0.99	0.62	0.74	0.98	0.63	0.64	0.85	0.90	0.93	0.87
	TEST_M2	0.885	0.97	0.49	0.63	0.95	0.89	0.97	0.99	1.00	0.99	0.95
	TEST_M5	0.200	0.42	0.56	0.18	0.33	0.38	0.06	0.01	0.02	0.01	0.03
	TEST_M6	0.136	0.35	0.55	0.05	0.06	0.09	0.01	0.01	0.26	0.00	0.00
JSSD300	TEST_ONE	0.943	0.97	0.88	0.93	0.95	0.91	0.95	0.92	0.98	1.00	0.94
	TEST_M1	0.675	0.94	0.87	0.72	0.98	0.47	0.24	0.89	0.80	0.80	0.50
	TEST_M2	0.829	0.99	0.29	0.80	0.99	0.66	0.76	1.00	1.00	0.96	1.00
	TEST_M5	0.868	0.83	0.74	0.77	0.95	0.86	0.89	0.99	0.95	0.86	0.84
	TEST_M6	0.593	0.60	0.52	0.03	0.82	0.84	0.75	0.78	0.97	0.07	0.55
JSSD256	TEST_ONE	0.921	1.00	0.79	0.84	0.98	0.82	0.87	1.00	1.00	1.00	0.92
	TEST_M1	0.646	0.91	0.50	0.53	0.93	0.48	0.17	0.93	0.93	0.71	0.45
	TEST_M2	0.863	0.98	0.64	0.77	0.99	0.74	0.83	0.99	0.99	0.93	0.91
	TEST_M5	0.864	0.82	0.82	0.63	0.95	0.79	0.90	0.99	0.99	0.93	0.95
	TEST_M6	0.585	0.66	0.49	0.00	0.87	0.81	0.70	0.73	0.73	0.97	0.59
JSSD224	TEST_ONE	0.930	0.97	0.81	0.91	1.00	0.82	0.87	1.00	1.00	1.00	0.92
	TEST_M1	0.644	0.93	0.53	0.53	0.91	0.53	0.18	0.88	0.66	0.84	0.45
	TEST_M2	0.853	0.98	0.47	0.73	0.99	0.75	0.87	1.00	0.90	0.89	0.95
	TEST_M5	0.822	0.82	0.68	0.46	0.99	0.75	0.90	0.95	0.92	0.79	0.95
	TEST_M6	0.572	0.66	0.50	0.00	0.77	0.68	0.76	0.72	0.95	0.11	0.56

of mAP 0.814, but for the rest, the mAP is about 0.436 in Fast-SSD, and 0.675, 0.646, 0.644 in the JSSD300, JSSD256, JSSD224 respectively. There is still room for improvement.

IV. CONCLUSION

In this paper, the SSD is mainly used for hand gesture detection and recognition. However, the SSD is limited when the hand gestures are collected more than 2 meters from the camera. In order to recognize the hand gestures at a longer distance, especially when the images collected more than 5 meters from the camera, the Joint SSD network is proposed to extend the recognition range. The principle of JSSD is that we have two cascaded SSDs for object detection named SSD1 and SSD2 connected by one middle Trigged layer as a bridge. The SSD1 is used for extracting the head-shoulder area. If the head-shoulder area is extracted successfully, the trigger layer is triggered. Then, the head-shoulder area feature will be fed into SSD2. The responsibility of SSD2 is to extract the hand area and classify the extracted area. The experimental results show that the JSSD network works good in recognizing the hand gestures at a longer distance under the bad lighting conditions. In the training process, the aspect ratio is picked at random, which is not an efficient and reasonable method. Therefore, the k-means is introduced to automatically the appropriate width to height ratio and aspect ratios of the bounding boxes.

Future work will focus on the speed acceleration of the model, we'll try to minimize the network and realize the real-time performance on the embedded systems. Besides, the mAP is also required to be improved.

V. ACKNOWLEDGEMENT

This work is a collaboration work of Central South University and the Chinese University of Hong Kong, Shenzhen. We would like to thank the Prof. Dongxiao Zhu in CUHKSZ for helpful discussions. Also, we would like to thank the Chunyue Xue at the Robotics and Artificial Intelligence Laboratory and volunteers at Central South University for data collection.

REFERENCES

[1] M. Martínez-Camarena, M. Oramas, and T. Tuytelaars, "Towards sign language recognition based on body parts relations," in *ICIP*, 2015.

[2] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *ICIP*, 2014.

[3] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *ICME*, 2015.

[4] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[6] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NIPS*, 2016.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.

[9] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv preprint arXiv:1710.09282*, 2017.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[11] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *arXiv preprint arXiv:1707.01083*, 2017.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2013.

[13] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[14] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[15] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and depth sensor," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14 991–15 015, 2016.