Long-Range Hand Gesture Recognition via Attention-based SSD Network

Liguang Zhou[1,2], Chenping Du[3], Zhenglong Sun[1,2], Tin Lun Lam[1,2,†], Yangsheng Xu[1,2]

***Abstract*— Hand gesture recognition plays an essential role in the human-robot interaction (HRI) field. Most previous research only studies hand gesture recognition in a short distance, which cannot be applied for interaction with mobile robots like unmanned aerial vehicles (UAVs) at a longer and safer distance. Therefore, we investigate the challenging long-range hand gesture recognition problem for the interaction between humans and UAVs. To this end, we propose a novel attention-based single shot multibox detector (SSD) model that incorporates both spatial and channel attention for hand gesture recognition. We notably extend the recognition distance from 1 meter to 7 meters through the proposed model without sacrificing speed. Besides, we present a long-range hand gesture (LRHG) dataset collected by the USB camera mounted on mobile robots. The hand gestures are collected at discrete distance levels from 1 meter to 7 meters, where most of the hand gestures are small and at low resolution. Experiments with the self-built LRHG dataset show our methods reach the surprising performance-boosting over the state-of-the-art method like the SSD network on both short-range (1 meter) and long-range (up to 7 meters) hand gesture recognition tasks.**

I. Introduction

The hand gesture is an intuitive way for humans to interact with robots. Recently, the deep learning based techniques surprisingly boost the hand gesture recognition algorithms in accuracy over the traditional hand-craft features [1-3]. Therefore, the hand gesture becomes one of the significant communication bridges in the human-robot interaction.

In the research community, the hand gesture recognition problem has been studied for decades. Traditional methods mainly focus on hand-craft features for hand gesture recognition, which shows promising results in the ideal conditions. However, these methods are not robust enough and suffer from sophisticated background clutter, occlusion, scale problem, various lighting conditions, and viewpoint problems.

Thanks to the fast development of the graphics processor unit (GPU) and the massive amount of the human-labeled dataset like the ImageNet, MS COCO, and the user-friendly programming library for deep neural network deployment like the Pytorch [4-6]. Many deep learning based hand gesture recognition systems have been proposed to alleviate the above-mentioned problems and enhance the robustness and efficiency of the traditional hand recognition algorithms. The hand gesture recognition system based on the SSD network shows that detection performance on short-range gestures is significantly outperforming the traditional methods [1, 3, 7]. Notably, the frames per second (FPS) of the one-stage object detector, like SSD, is faster than models with two-stage detectors. Hence, the SSD [7] network is suitable for real-time hand gesture recognition. However, when the SSD is applied to detect long-range hand gestures, it leads to poor performance, especially when the interaction distance between the robot and human is more than 5 meters. Most existing hand gesture recognition systems are limited on the recognition distances [1, 8-11].

To recognize the long-range hand gestures, the Joint SSD [3] is proposed, where the first SSD is designed to detect the head-shoulder area, and the second SSD is designed for detection and classification of hand gestures. However, the bounding box of the hand-shoulder area is required for the first SSD, making it more labor-intensive and less efficient. Moreover, the Joint SSD performs worse on the short-range hand gesture recognition compared with the SSD. Besides, it demands more GPU memory to deploy the model.

In many industrial applications [12-14], for humans to interact with the UAVs, a safer and longer distance is required, and a more lightweight and efficient model is in demand. To this end, we present an attention-based SSD for recognizing hand gestures at both short-range and long-range. Expressly, the convolutional block attention module (CBAM) that combines both the spatial and channel attention are incorporated into the SSD, namely CBAM-SSD, which significantly prolongs the recognition distance from only 1 meter to 7 meters.

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

- We release a dataset, namely **Long-range Hand Gesture (LRHG) Dataset** *, which features the small hand gestures collected at long-range by a lightweight USB camera, enabling the interaction between human and robots at multiple ranges and long-range via hand gestures.
- To solve the novel long-range hand gesture recognition problem, we incorporate both the spatial and channel attention mechanisms into the SSD network with a CBAM module, namely CBMA-SSD. The performance of the long-range hand gesture recognition has been tremendously boosted. Experiment results show the

† Corresponding Author: Tin Lun Lam (tllam@cuhk.edu.cn).

* This paper was supported by funding 2019-INT007 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society and the Shenzhen Science and Technology Innovation Commission, fundamental research grant JCYJ20170818104502599.

[1] The Chinese University of Hong Kong, Shenzhen
[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3] Sun Yat-sen University

*https://sites.google.com/view/lrhgd/dataset

proposed CBAM-SSD notably outperforms state-of-the-art methods and recognizes hand gestures collected up to 7 meters.

The rest of the paper is structured as follows. Section II introduces the related works. Section III describes the CBAM based SSD detector for more reliable hand gesture detection, which can be used for human-robot interaction at a longer and safer distance. Section IV shows the experimental results and demonstrates that the proposed CBAM-SSD notably increases long-range hand gesture recognition performance and improves short-range hand gesture recognition. Finally, the conclusion is summarized in Section V.

II. Related Works

The literature on hand gesture detection can be divided into two categories, data gloves-based method (contact) and non-contact hand gesture detection [15]. In this paper, we mainly discuss the non-contact hand gesture recognition methods. Moreover, we discuss the object detectors for object detection and attention mechanisms for network performance enhancement.

A. Non-Contact Hand Gesture Recognition

In non-contact hand gesture detection, the wireless technology based method, like the Two-Antenna Doppler Radars, has a limited sensing range of 0.5 meters and is applied to detect the hand gestures based on the deep convolutional neural networks [16]. Moreover, there are many works about vision-based methods because of their natural and non-contact features. Xu et al. [9] propose an appearance-based hand gesture recognition system that utilizes color and depth images. Marin et al. [1] present a multi-class SVM classifier to train the dataset based on the ad-hoc feature set, where the position and orientation of the fingertips are calculated. Generally, sign language consists of two parts, one is hand posture, represented by the position and configuration of fingers, and the other is the hand gesture, which means the moving trajectory of the hand [17]. However, the algorithm demands the depth information from Kinect and the hand points, and features from the Leap Motion to boost the recognition accuracy [18]. Meanwhile, the sensing distance of the Leap Motion is relatively small. Moreover, the hand-crafted feature descriptors are complex and require a lot of effort to design. Huang et al. [2] integrate the multi-channel information like the color images, depth images, and body skeleton images obtained from the Kinect as input to 3D CNNs. Still, they only investigate short-range hand gesture recognition. In summary, the above models are limited to the short-range hand gesture recognition tasks. To make long-range hand gesture recognition possible, the Joint-SSD [3] has been proposed, but the performance for short distance hand gesture detection is decreased [3]. Moreover, the Joint SSD requires two bounding boxes, one is the head-shoulder area, and the other is hand gestures, which makes the training process more complicated. Besides, it introduces an extra amount of work for the labeling of the head-shoulder area. Mazhar et al. [11] propose the body skeleton based hand gesture recognition system for human-robot interaction, but with a low fps (5.2fps) and limited interaction range (4m).

B. Deep ConvNet Object Detectors

In the object detection area, many deep learning based models like R-CNN [19], Fast R-CNN [20], Faster R-CNN [21] have been developed and shows dramatic improvements in terms of accuracy. The detection process of these methods includes two steps: region proposal and image classification. Also, some improved networks do not rely on region proposals such as SSD [7], and YOLO [22, 23]. In general, models without region proposals have faster speed than models with region proposals. Compared with the R-CNN, Fast R-CNN, Faster R-CNN, SSD features a faster speed and more lightweight network design.

C. Attention Mechanism

Recently, the attention mechanism shows the potential and effective capabilities in the NLP and computer vision areas. Self-attention is used for machine translation with a transformer [24]. Non-local computes the weighted mean of all pixels as a typical filtering algorithm [25] and Wang et al. [26] introduced the non-local network for video classification recently. However, the local information that is essential for small hand gesture recognition is not considered in non-local attention. SENet [27] assumes that each channel's importance in the feature map of the deep neural network varies. It utilizes a global pooling method to measure the importance of each channel in the feature map. The spatial information is missed when using the SENet. Woo et al. propose the CBAM [28] module, which combines both the spatial and channel information of feature map, is more suitable for recognition of hand gestures.

III. Methodology

In this section, we introduce the fundamental object detector SSD for hand gesture recognition that performs excellently on short-range hand gesture recognition. But it is not capable of handling the task of long-range hand gesture recognition. Thus, to improve the performance of long-range hand gesture recognition, we improve the SSD by incorporating the spatial and channel attention module with the CBAM, convolutional block attention module (CBAM), namely the CBAM-SSD network.

A. Single Shot Multibox Detector (SSD)

1) Basic Network: SSD is a cutting-edge deep neural network for object detection. The architecture of the SSD model for hand gesture recognition is depicted in Fig. 1, where the SSD is the basic object detection model for hand gesture detection. For the feature extraction part, the backbone network is the VGG16 network. Unlike the Faster-RCNN, when localizing the bounding box, the selective search region proposal method is used to generate the potential bounding boxes. In the SSD, the anchors are applied to generate bounding box proposals.

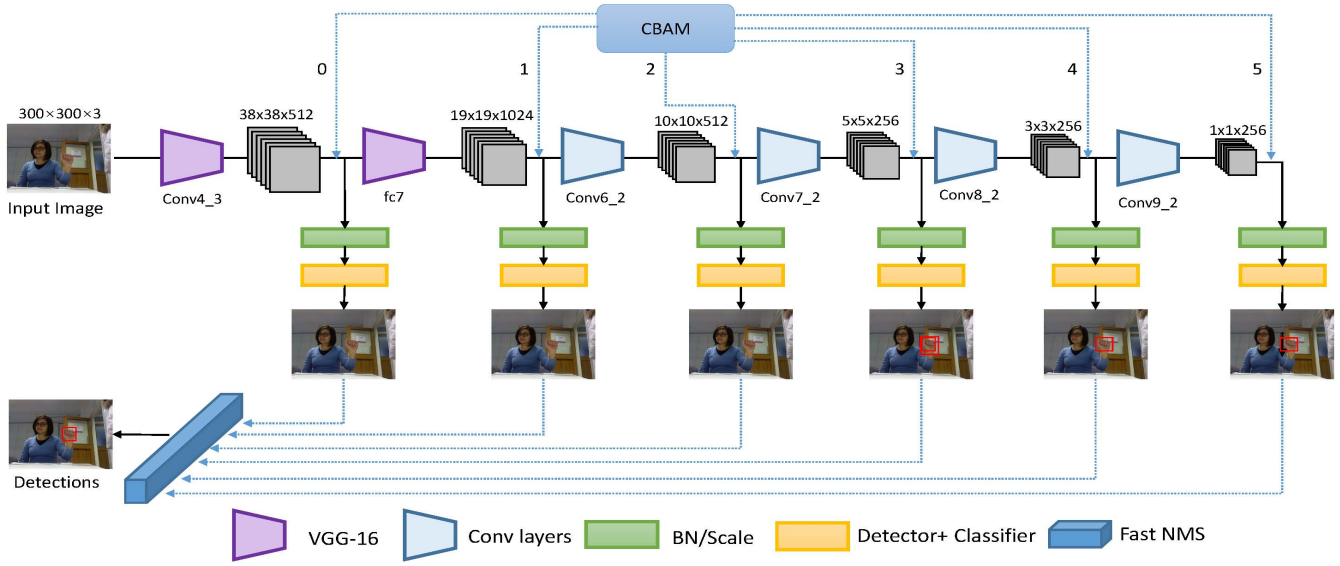


Fig. 1. The network structure of the proposed CBAM-SSD model is shown in the Figure. We connect each output feature map with a CBAM module individually ranging from CBAM0-SSD to CBAM5-SSD, to better capture the feature representations of both the short-range and long-range hand gestures.

2) loss function: The loss function of SSD can be found in literature [7]. The overall objective loss function is a combination of the weighted sum of the localization loss and the confidence loss:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)) \quad (1)$$

where N is the number of default boxes matched (if N=0, L=0), and the weights term α is set to 1. L_{loc} is the localization loss function of smooth L1 between the predicted box (l) and ground truth box (g):

$$\begin{aligned} L_{loc}(x,l,g) &= \sum_{i\in Pos, m\in\{cx,cy,w,h\}} \sum x_{ij}^{k}\,\mathrm{smooth}_{L1}\left(l_i^m - \hat{g}_j^m\right) \\ \hat{g}_j^{cx} &= \left(g_j^{cx} - d_i^{cx}\right)/d_i^w \quad \hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^h \\ \hat{g}_j^{w} &= \log\left(\frac{g_j^w}{d_i^w}\right) \quad \hat{g}_j^{h} = \log\left(\frac{g_j^h}{d_i^h}\right) \end{aligned} \quad (2)$$

$$\mathrm{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (3)$$

where $X_{ij}^p = \{0,1\}$ is the indicator for matching the i-th default box to the j-th ground truth box in category p. We know $\sum_i X_{ij}^p \geq 1$ in the matching strategy. The (cx,cy) is the center of the bounding box (d), and w, h are the width and height of the bounding box. The confidence loss L_{conf} is the Softmax Loss, c represents the input confidence.

$$\begin{aligned} L_{\mathrm{conf}}(x,c) &= -\sum_{i\in Pos}^{N} x_{ij}^p \log\left(\hat{c}_i^p\right) - \sum_{i\in Neg} \log\left(\hat{c}_i^0\right) \\ \text{where} \quad \hat{c}_i^p &= \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \end{aligned} \quad (4)$$

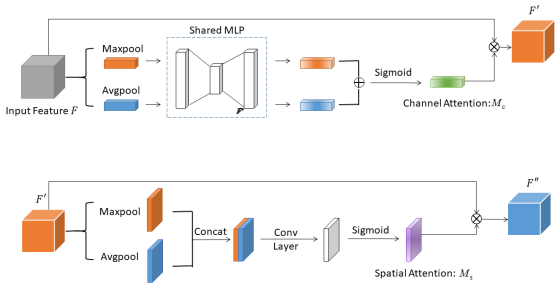


Fig. 2. The figure shows the network structure of the CBAM module, where the top flow shows the channel attention module and the bottom flow shows the spatial attention module. The Input feature of the CBAM module is the intermediate feature map of the neural network, and the output feature will be connected to the next layer of the network. With the CBAM module, better feature representations of the input object can be learned.

B. Convolutional Block Attention Module (CBAM)

The architecture of the CBAM is depicted in Fig. 2. There are two essential blocks in CBAM [28]. One is channel attention, which is tailored for exploiting the inter-channel relationship of features. The other is spatial attention, which is designed for capturing the spatial relations of feature maps. The mathematical expression of Channel Attention is:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (5)$$

where σ denotes the Sigmoid function and f^{7x7} represents a convolution operation with the filter size of 7*7. MLP is the multilayer perceptron of the three layer fc network share by features of max pooling and average pooling. The AvgPool and MaxPool represent the operations of average pooling and max pooling, respectively. Similarly, the mathematical

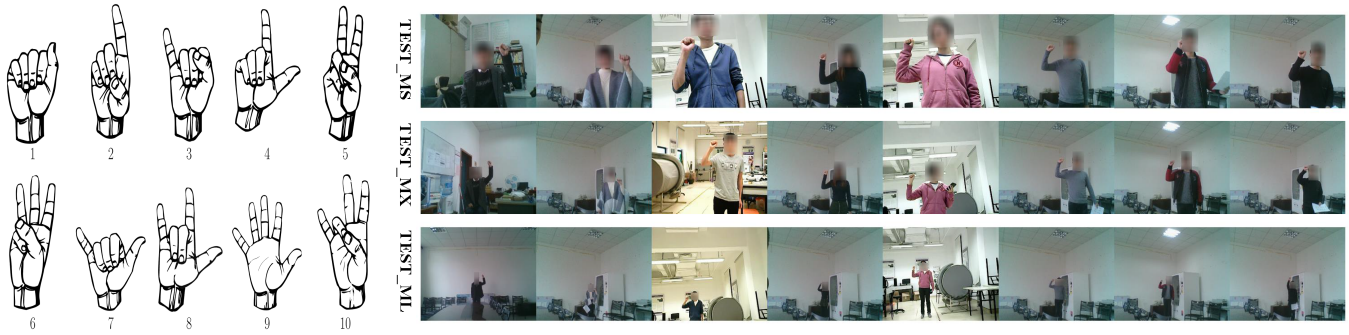


Fig. 3. The left figure shows the 10 hand types of the LGRH Dataset. The right figure shows the example images from test sets of the LRHG dataset.

expression of Spatial Attention is:

$$M_s(F) = \sigma(f^{7x7}([AvgPool(F); MaxPool(F)])) \quad (6)$$

Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, the channel attention map will be computed first as $M_c((F) \in \mathbb{R}^{C\times 1\times 1}$ and F' is the intermediate feature for spatial attention calculation, where F'' is obtained with $M_s(F') \in \mathbb{R}^{1\times H\times W}$ as the final output of the CBAM. The CBAM pipeline is summarized as follows, where $\otimes$ represents the element-wise multiplication:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \quad (7)$$

C. CBAM-SSD

SSD only recognizes the hand gestures collected less than 2m from the robot. Besides, as the distance increases, the performance of SSD becomes unreliable and dramatically decreases. In the interest of recognizing the hand gestures at a longer distance, the CBAM is introduced to improve the recognition performance of the SSD, based on the assumption that attention can help the algorithm to focus on the target objects during the training process automatically. Therefore, we incorporate the SSD and CBAM, namely CBAM-SSD for hand gesture recognition depicts in Fig. 1. One thing that can be noticed is that there are six distinct feature maps for the box proposal modules. To better capture the feature maps with the help of the attention mechanism automatically, we attach the CBAM module to each one of these six modules individually, ranging from the CBAM0 to CBAM5. We have conducted the comparison experiments by adding the CBAM0 to CBAM5 into the SSD model individually, ranging from the CBAM0-SSD to CBAM5-SSD. Surprisingly, the proposed method can increase the performance of hang gesture recognition both in short-range and long-distance.

IV. Experiment Results

A. Experimental Settings

In this paper, we aim to increase the recognition accuracy for the long-range hand gestures, therefore enabling human-robot interaction at a safer and longer distance via hand gestures. To this end, we evaluate the proposed method on the long-range hand gesture dataset. We will introduce the implementation details and training procedures, and different experiment settings.

TABLE I
LRHG Dataset

LRHG Dataset					
Multi-Range				Single Range	
Training	Test			Tiny Hand	Median Hand
Train_MX	Test_MX	Test_MS	Test_ML		
3024	1296	543	504	1695	1796

1) Implementation Details: For the models, we deploy in the experiment, the optimizer used is the Stochastic Gradient Descent (SGD) with an initial learning rate 0.0001, the momentum of 0.9, weight decay 0.0005. The backbone network is VGG16 pretrained on the ImageNet [29]. The total number of iterations is 20000. The model evaluation criterion is MultiBoxLoss. The batch size of the training data is 32. During the learning process, both the VGG16 and the SSD parts of the model are updated based on the SGD optimizer. The speed of the model is tested on a single GTX 1080 Ti.

2) LRHG Dataset: In LRHG Dataset, there are 10 different types of hand gestures collected from 8 different people at discrete distance levels, such as 1m,2m,3m as the short-range and 5m,6m,7m as long-range as Fig. 3 presented. The images are captured by a lightweight USB camera at a speed of 25fps in the lab environment without strict illumination constraints. The size of the captured images is 640×480 pixels, and the type of images is RGB color. Each gesture is performed by 8 people at discrete distance levels. The collected dataset is **Long Range Hand Gesture Recognition (LRHG) Dataset**. To make the dataset more challenging, we set the hand gestures performed by people with different scales, viewpoints, and in-plane rotation.

As Tab. I shows, the LRHG dataset has 4320 images in total. The LRHG dataset is split into the training dataset TRAIN_MX with 3024 images and test dataset TEST_MX with 1296 images at a ratio of 7:3. There are two extra test datasets taken from the LRHG dataset. The TEST_MX includes hand gestures collected at both short-range and long-range. The TEST_MS, in which the images are from the Test_MX and are mostly collected at short-range from the camera. The Test_ML, in which the images are from the

Test_MX and are mostly collected at long-range from the camera. Moreover, each dataset contains all types of hand gestures.

To study the median-range hand gesture recognition problem that hand size lies between 32x32 and 64x64, one subset from the LRHG dataset, namely the Median Hand Dataset, is built. Similarly, to investigate the long-range hand gesture recognition task, where the hand area is less than 32x32, one subset from the LRHG dataset, namely Tiny Hand Dataset, is built.

As histograms presented in Fig. 4, compared with the existing dataset [1, 17], our dataset features 5x smaller hand gestures in the area and more than 10x smaller on average, which indicates hand gestures are collected at a much longer distance, and more hand gestures have been collected at multiple discrete distances, thus achieving the long-range human-robot interaction using the RGB camera mounted on mobile robots.

B. Comparison with Hand Gesture Recognition Systems

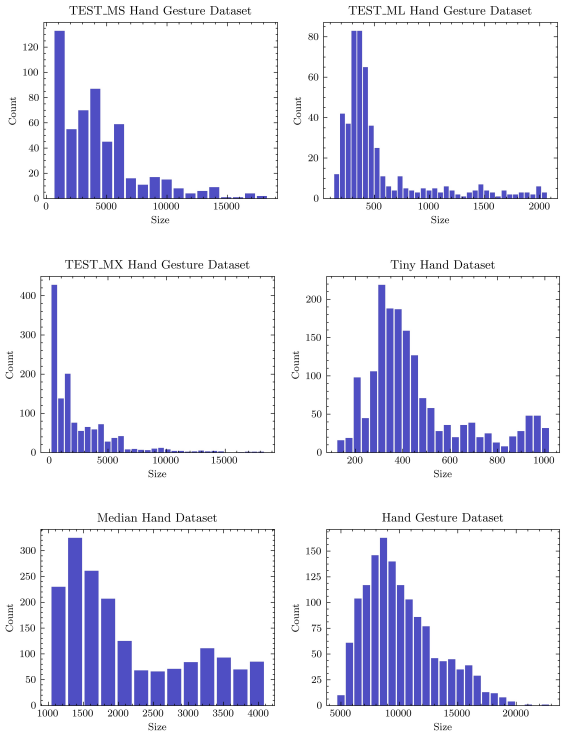


Fig. 4. Histograms of LRHG Dataset and previous dataset, where the x-axis is the size of hand gesture in pixels, and the y-axis is the number of images. We separate LRHG into the training set and test set, namely TRAIN_MX and TEST_MX, where two extract test sets, TEST_MS and TEST_ML, are taken from Test_MX. Besides, we also separate two datasets, namely Tiny Hand Dataset, Median Hand Dataset, to test the small hand and median hand recognition ability of the proposed method. The size of a tiny hand is less than 1000 pixels and mostly are less than 512 pixels. The size of median hands is between 1000 and 4000 pixels, which are much smaller than the Hand Gesture Recognition Dataset in [1].

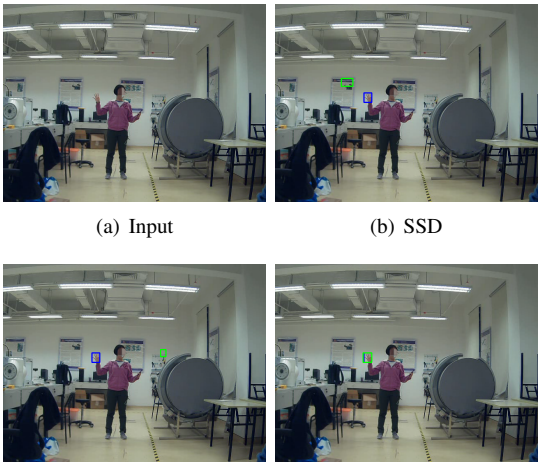
As Tab. II shows that the proposed hand gesture recognition system has three benefits compared to other hand gesture recognition systems. The first one is that the proposed system features the recognition of small hand gestures and recognizes the hand gesture up to 7 meters, while most of the previous systems are mainly interact with hand gestures within 1 meter. Compare with the Skeleton+CNN that reaches up to 4 meters, which is limited for interaction with mobile robots. Besides, the Kinect V2 sensor is required, which is not lightweight and has a limited depth-sensing ability (4.5 meters). In contrast, our system can interact with hand gestures up to 7 meters, which is long enough and safer for interaction with UAVs. The third benefit is the speed of our system reaches 28.3 fps, which is 5x faster than the Skeleton+CNN system.

TABLE III
COMPARISON ON THE MAP FOR DIFFERENT CONFIGURATION OF THE CBAM BASED SSD

	TEST_ML	TEST_MS	TEST_MX
SSD	32.1	96.1	68.7
CBAM0-SSD	91.0	98.3	93.6
CBAM1-SSD	91.2	**99.1**	93.7
CBAM2-SSD	90.9	97.4	93.4
CBAM3-SSD	93.3	97.9	94.0
CBAM4-SSD	92.8	97.5	93.2
CBAM5-SSD	**93.7**	**98.1**	**94.5**

The qualitative samples are displayed in Fig. 5. The input image is captured by a small UAV at 5 meters away from a person, where the hand gesture is relatively small, as we observed (17*19 in pixels). The proposed CBAM-SSD accurately localizes the small hand gesture while the SSD and JointSSD fail, showing that the proposed method can produce high-quality hand recognition results.

C. Experiment Results



(a) Input (b) SSD

(c) Joint SSD (d) CBAM-SSD

Fig. 5. Qualitative results comparison with various methods is shown in this figure. The input image is captured by a UAV at 5 meters away from the person, where the hand gesture is in a relatively small resolution (17*19 in pixels). The green is the bounding box predicted, while the blue is the ground truth bounding box. The proposed CBAM-SSD network can accurately predict the small hand gesture's bounding box at the long-range, while the Joint SSD and SSD network fail.

To evaluate how the performance is affected by the position of the CBAM in the SSD network. We have conducted

TABLE II

SYSTEM CONFIGURATIONS AND PERFORMANCE COMPARISON OF PROPOSED HAND GESTURE RECOGNITION SYSTEM WITH OTHERS

Methods	Recognition Range	Accuracy	mAP	Speed	Person number	Data Type	Sensors
ANNs [8]	0m	99.0	-	Real-time	-	-	Glove
SRC+Joint Feature [9]	1m	93.8	-	Slow	-	RGB+RGB-D	Kinect
multi-class SVM [1]	1m	91.3	-	Slow	14	RGB-D	Kinect+Leap Motion
HMM [10]	1m	87.5	-	-	-	RGB-D	Kinect + Glove
Joint SSD [3]	5m	-	83.6	13.5 fps	8	RGB	USB Camera
Skeleton+CNN [11]	4m	95.7	-	5.2 fps	-	RGB-D	Kinect V2
Ours	**7m**	-	94.5	28.6 fps	8	RGB	USB Camera

TABLE IV

COMPARISON ON THE TOTAL PARAMETERS AND MAP WITH THE STATE-OF-THE-ART METHODS

	Params	TEST_ML	TEST_MS	TEST_MX
SSD [7]	24.95M	31.7	95.4	68.7
Joint SSD [3]	49.90M	83.4	91.0	83.6
CBAM5-SSD	24.96M	**93.7**	**98.1**	**94.5**

ablation studies on the different configurations of the CBAM based SSD in Tab. III. The experiment result shows that as the CBAM is attached to the deeper feature maps, the detection performance will be better, based on the assumption that the more deep feature representations of the network are, the better representations about the hand gestures can be generated for detection and classification.

Based on the LRHG Dataset hand gesture dataset, we conduct comparison experiments with other state-of-the-art methods. In comparison with state of the art, Tab. IV shows that CBAM5-SSD significantly boosts the mAP of SSD by 60.0 and Joint SSD by 10.3 on the Test_ML dataset, and it reaches 93.7 mAP. Besides, the CBAM5-SSD outperforms the Joint SSD by 7.1 on the TEST_MS and 10.9 on the TEST_MX dataset, showing the superior performance of CBAM5-SSD over the SSD and Joint SSD. These results demonstrate that CBAM-SSD is successful in hand gesture recognition at both short and long distances with competitive performance. This considerable improvement in the effectiveness of the CBAM5-SSD over state-of-the-art surprises our reasonable assumption of attention would be a performance increment for long-range hand gesture recognition algorithm.

To evaluate the proposed method's performance on the long-range hand gesture recognition, we split the Tiny Hand Dataset into training and test dataset as 8:2. The experiment result is shown as Tab. V, where G1 to G10 represents the gesture 1 to 10 as Fig. 3 shows. The proposed algorithm reaches 97.7 mAP, surpasses the SSD with 25.3 on the Tiny Hand Dataset.

Similarly, to evaluate the proposed method's performance on median-range hand gesture recognition, we split the Median Hand Dataset into training and test dataset as 8:2. The experiment result is listed in Tab. VI. The proposed algorithm reaches 98.5 mAP, surpasses the SSD with 4.5 on the Median Hand Dataset. Both experiments validate the effectiveness of the proposed method in recognizing both the median-range and long-range hand gestures.

TABLE V

PERFORMANCE OF THE PROPOSED METHOD ON TINY HAND DATASET

	SSD	CBAM5-SSD
G1	73.4	100.0
G2	70.9	89.0
G3	41.7	88.5
G4	76.5	100.0
G5	89.4	100.0
G6	57.1	99.4
G7	64.2	100.0
G8	64.2	99.6
G9	72.1	100.0
G10	88.7	100.0
mAP	72.4	**97.7**

TABLE VI

PERFORMANCE OF THE PROPOSED METHOD ON THE MEDIAN HAND DATASET

	SSD	CBAM5-SSD
G1	96.8	99.6
G2	89.0	98.4
G3	82.0	89.9
G4	96.0	99.4
G5	94.2	99.3
G6	92.0	99.8
G7	99.5	100.0
G8	98.6	100.0
G9	99.4	100.0
G10	92.7	100.0
mAP	94.0	**98.5**

V. CONCLUSION

The SSD is mainly applicable for hand gesture recognition in a short-range. i.e., when the hand gestures are collected at a long-range, the recognition performance of the SSD is suffering from a dramatic decrease. To recognize the hand gestures at a long distance, we incorporate an attention mechanism with the SSD network based on the assumption that the attention mechanism helps the model to focus on the region of interest in an unsupervised manner during training. Concretely, we attach the CBAM to the six feature maps of the SSD, ranging from the CBAM0-SSD to CBAM5-SSD. We test the SSD and CBAMx-SSD on both self-built short-range and long-range hand gesture datasets. Experiment results show CBAM-SSD dramatically improves the recognition performance at both short and long distances over the state-of-the-art methods. In the future, the proposed system will be applied to conduct the human-robot interaction task with the indoor service robots or UAVs at a longer distance.

REFERENCES

[1] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *2014 IEEE International conference on image processing (ICIP)*. IEEE, 2014, pp. 1565–1569.

[2] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *2015 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2015, pp. 1–6.

[3] C. Yi, L. Zhou, Z. Wang, Z. Sun, and C. Tan, "Long-range hand gesture recognition with joint ssd network," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2018, pp. 1959–1963.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[6] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] P. Neto, D. Pereira, J. N. Pires, and A. P. Moreira, "Real-time and continuous hand gesture spotting: An approach based on artificial neural networks," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 178–183.

[9] D. Xu, Y.-L. Chen, X. Wu, W. Feng, H. Qian, and Y. Xu, "A novel hand posture recognition system based on sparse representation using color and depth images," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3765–3770.

[10] F. Chen, Q. Zhong, F. Cannella, K. Sekiyama, and T. Fukuda, "Hand gesture modeling and recognition for human and robot interactive assembly using hidden markov models," *International Journal of Advanced Robotic Systems*, vol. 12, no. 4, p. 48, 2015.

[11] O. Mazhar, S. Ramdani, B. Navarro, R. Passama, and A. Cherubini, "Towards real-time physical human-robot interaction using skeleton information and hand gestures," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–6.

[12] Y. Wang and B. Ren, "Quadrotor-enabled autonomous parking occupancy detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 8287–8292.

[13] Y. Wang, Y. Wang, and B. Ren, "Energy saving quadrotor control for field inspections," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020.

[14] Y. Wang, Q. Sun, T. Berger, and W. Qi, "A drive-through recharging strategy for a quadrotor," in *2021 IEEE International Conference on Robotics and Automation*. IEEE, 2021.

[15] L. Yun, Z. Lifeng, and Z. Shujun, "A hand gesture recognition method based on multi-feature fusion and template matching," *Procedia Engineering*, vol. 29, pp. 1678–1684, 2012.

[16] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna doppler radar with deep convolutional neural networks," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3041–3048, 2019.

[17] M. Martínez-Camarena, M. Oramas, and T. Tuytelaars, "Towards sign language recognition based on body parts relations," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 2454–2458.

[18] 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.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[20] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[25] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 60–65.

[26] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[28] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.