Hand Gesture Recognition Based on Faster-RCNN Deep Learning

Xiaoguang Yu[1], Yafei Yuan[2*]
[1] Department of Art, West Anhui University, Luan 237012, China.
[2] Department of Electronic Engineering, Fudan University, Shanghai, 200433, China.

[*] Corresponding author. Email: yafeiyuan163@163.com
Manuscript submitted August 30, 2018; accepted October 31, 2018.
doi: 10.17706/jcp.14.2.101-110

Abstract: The hand gesture recognition is an effective way to apply human computer interaction which should aim at good recognition accuracy and high speed. To solve the problem of hand gesture recognition in complex scenes, we propose a new approach for hand gesture recognition which is based on the faster regional convolutional neural network (Faster-RCNN) deep learning algorithm with five layers of neural network. A benchmark database with gestures is used, several general characters under the complex environmental background are chosen to as the processing object. The model is established with a certain number samples for training. The results of model checking show that the algorithm can identify the gesture recognition categories effectively, quickly and accurately with low computational cost. The accuracy can reach highly up to 99.2% which is great significance for human computer interaction application.

Keyword: Human computer interaction, hand gesture recognition, image processing, deep learning, faster regional convolutional neural network.

1. Introduction

With the rapid development of computer technology and intelligent development, non-contact gesture recognition plays an important role in human computer interaction (HCI) [1]-[3]. The hand gesture recognition system can be used for an effective and natural human-machine interaction. What more, the gesture recognition based on vision is widely applied in artificial intelligence, multimedia, virtual reality and natural language communication [4]-[7].

However, the gesture recognition based on image processing algorithm was not applied and developed widely duo to its poor real time capable, low accuracy recognition and complex algorithm. In recently years, as the application of graphics processor (GPU) to artificial intelligence (AI), the gesture recognition based on machine learning has been developed rapidly in wide application. The machine learning algorithms such as local orientation histogram, support vector machine (SVM)[8], neural network and elastic graph matching are widely used in gesture recognition system [9]-[11]. The neural network owns the good characteristic learning ability. It does not need manual feature setting during simulating human learning process and can carry out training the gesture sample to form a network classification recognition map [12], [13]. Deep learning, a relatively recent approach to machine learning, which involves neural networks with more than one hidden layer. It can acquire the characteristics of the learning object easily and accurately under the complex object and exhibit the superior performance in face recognition, speech recognition and Natural Language Processing (NLP) [14]-[16]. The Fast RCNN and Faster RCNN make further optimization

for the object detection. Compare with the traditional RCNN, the Fast/Faster RCNN which is initialized with discriminative pre-training for ImageNet classification uses convolutional layers to extract region features followed by a region wise multilayer perceptron (MLP) for classification.

In this paper, the gesture recognition method based on the faster region full convolution neural network is proposed to deal with hand gesture recognition in complex environment. The experiment results reveal that the Faster RCNN can overcome the interference signals in complex background and improve the accuracy and processing speed of gesture recognition.

2. Related Work

Generally, three basic processing stages consists the vision-based hand gesture recognition system including hand segmentation, gesture modelling, and finally gesture classification. The hand gesture recognition system was divided into the training phase and the testing phase. The training phase includes the hand segmentation and gesture modelling processing stages, while the testing phase includes all three processing stages.

The main purpose of hand segmentation is to detect the hand regions from the hand gesture sequence, separate them from the complex backgrounds and provide the effective input information source for the following training phase. In the stage of hand gesture analysis, hand postures as well as motion patterns are calculated from the hand gesture frame sequence. Training is used to acquire an effective recognition model; recognition output is based on the model that has been trained to identify the gesture categories of input data. Then the hand gesture model is created accordingly. The final stage is gesture classification by the built model of the hand gesture. The gesture classification has to be both fast and accurate. A multilayer perceptron (MLP) network is used to recognize the hand posture, and the feature vector is input to this network. An MLP is a standard method for nonlinear classification and can approximate any continuous function to an arbitrary accuracy. However, the network's performance within a given training period is dependent on its structure. A small network may have limited processing capability, while a large one may be too slow and may have redundant connections. Thus, the same evolutionary algorithm is also used to select the optimal layout for the neural network.

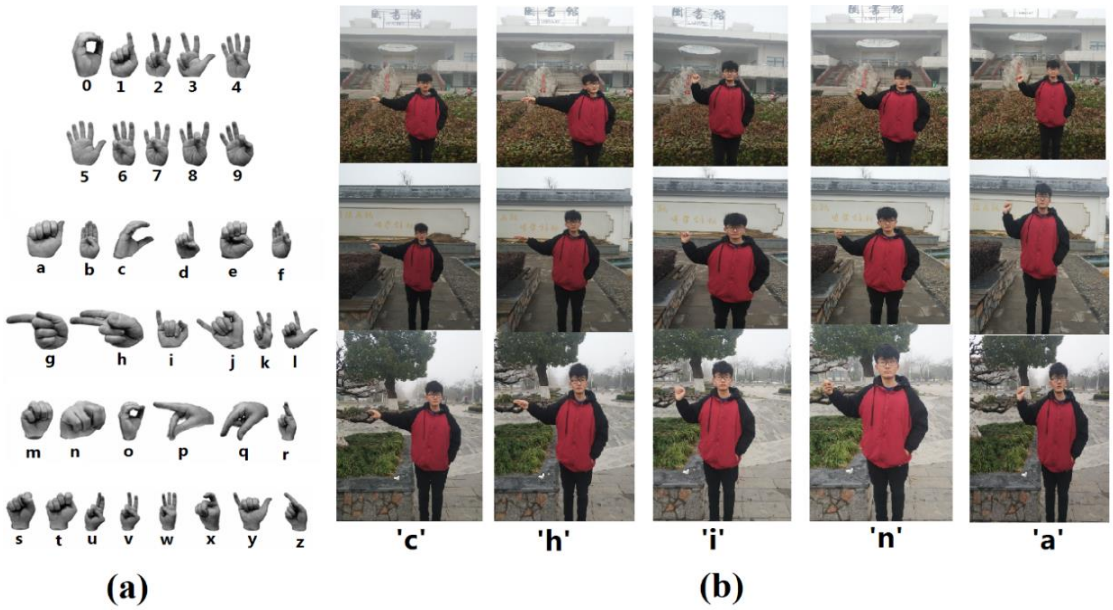

Fig. 1. The graphs of hand gestures information. (a) Standard library of hand gestures; (b) Five hand gestures graphs in different backgrounds.

The standard hand gesture database is key important in human-machine interaction system. The Fig. 1(a) shows the standard numbers and letters with hand gestures database. Note that some gestures are rather difficult to distinguish. For example "a" and "e", "d" and "l", "m" and "n" or "i" and "j". In this work, the characters of "c", "h", "i", "n" "a" are chosen as the study objects which was shown in Fig. 1(b). Each hand gesture sample is obtained under three different complex background which is aim at proving the applicability and reliability of the hand gesture recognition system.

The hand segmentation detection plays a preprocessing role in gesture recognition system which is regarded as the first step towards process the input gestures. Here, the input and output are images, these images are transformed to the binary form in order to prepare original image for the region segmentation. The binary images are obtained by threshold the original images at 0.5 gray level. The purpose of segmentation technique is mainly to divide the special domain on which the image is defined. An algorithm to segment hand-occupied regions in the binary images is then implemented. The hand segmentation algorithm tracks the boundary of the white pixels in the images, and extracts pixels in form of a rectangular bounding box containing the segmented hand then obtain the meaningful regions from the original image. The meaningful region may be a complete object or may be a part of it.

In this work, the skin color segmentation is applied to carry out the preprocessing step in gesture recognition system under the complex background. The skin color has been utilized by several approaches for hand detection. A major decision towards providing a model of skin color is the selection of the color space to be employed. Several color spaces have been proposed including RGB, normalized RGB, HSV, YCrCb, YUV, etc. HSV is an effective color space method for skin color segmentation. HSV is an effective color space method for skin color segmentation. As the image resources are BGR color space, the color conversion is needed for HSV color space by employing following expression [17].

$$V = \max(B, G, R) \tag{1}$$

$$S = \frac{V - \min(B, G, R)}{V} \tag{2}$$

$$H = \begin{cases} 240° + 60° * \dfrac{V - G}{V - \min(B, G, R)}, V = B \\ 120° + 60° * \dfrac{V - R}{V - \min(B, G, R)}, V = G \\ 60° * \dfrac{V - B}{V - \min(B, G, R)}, V = R \end{cases} \tag{3}$$

where H is the shades of color ($0 \backsim 360°$), S is the saturation of color, V is the brightness of color. The *B, G, R* are all normalized values and the value is [0, 1]. This method owns the advantages of simple rule, fast speed and highly accuracy, so as to detect the meaningful region gesture quickly and accurately. After the hand segmentation detection, the neural network is trained with the obtained meaningful region data by machine deep learning [18].

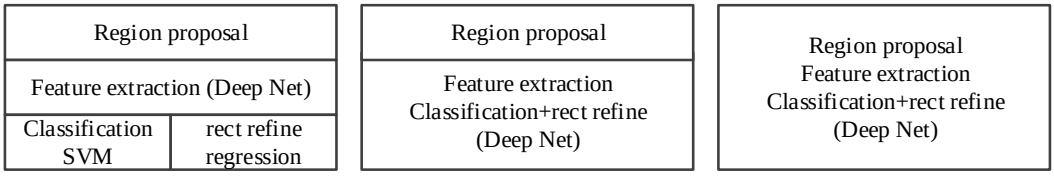

Fig. 2. The diagram of algorithm.

The Faster-RCNN algorithm is proposed to the gesture recognition system due to its highly efficient and accurate. The Faster-RCNN is evolved from RCNN and Fast-RCNN [19], [20]. The relation of Faster-RCNN, Fast-RCNN and RCNN algorithms is shown in Fig. 2.

The successful RCNN [21] algorithm applies high-capacity convolutional neural networks to extract a fixed-length feature vector from each region which is fed to a set of class-specific linear SVM. It firstly was trained the network by supervision for image classification with abundant data and then fine-tunes the network for detection where data is scarce. In fact, it only can be considered a hybrid of traditional. Although this method has many advantages over the traditional algorithm, there are also shortcomings such as cumbersome training steps, long time consuming, large hard disk occupying speed and slow speed.

Fast-RCNN and Faster-RCNN make further evolution on the pipeline of object detection. Following the pioneering RCNN, Fast/Faster RCNN uses convolutional layers, initialized with discriminative pre-training for ImageNet classification, to extract region-independent features followed by a region wise multilayer perceptron (MLP) for classification. Besides, they jointly optimize a soft max classifier and bounding-box regressor, rather than training a soft max classifier, SVMs, and regressor in three separate stages. The Fast-RCNN algorithm uses a selective search algorithm to extract a specific number of advice windows in the image by using a selective search algorithm, then the whole image is input into the CNN for feature extraction. The proposed window is mapped to the last layer convolution feature map of CNN which is made through the ROI pool layer. The characteristic map of the fixed size is produced by windows. Finally, the classification probability and border regression training are combined using the detection classification probability and the detection border regression algorithm. Compared with RCNN, the Fast-RCNN is added to the ROI pool layer after the final convolution; the loss function uses the multitask loss function and the border return is directly added to the CNN network to train and save the computing resources. Fast-RCNN, the improved algorithm has many advantages. After normalization of the whole image, it is directly sent into the CNN, and the suggestion box information is added to the feature graph of the final convolution layer, so that the CNN operation is shared with avoiding the waste of computing power and raising the speed of the test time in the training process. Only one image is needed to be sent to the network, each image extracts the CNN feature and the proposed window at one time. The training data is directly into the loss layer in the memory of the graphics processing unit (GPU). The first layers of the candidate region do not need to repeat the calculation and no longer need to store a large amount of data on the hard disk. According to the problem of large hard disk space and slow speed in training, the algorithm unify the classification and position regression with the depth network, without the consumption of additional computing resources, so as to improve the speed and timeliness in the recognition system.

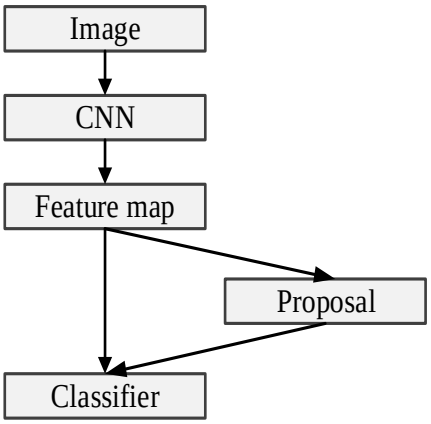

Fig. 3. The flow chart of fast-RCNN algorithm.

The Faster-RCNN algorithm directly inputs the entire detection image into the CNN for feature extraction, then outputs a proposed window with Region Proposal Network (RPN). Each image generates a recommended window for the characteristics, then maps the proposed window to the final convolution feature graph of CNN by generating a fixed size for each ROI through the ROI pool layer. Finally, the classification probability and border regression training are combined using the detection classification probability and the detection border regression algorithm. Faster-RCNN compared with Fast-RCNN, using RPN instead of the original selective search method produces a suggestion window. The CNN of the proposed window and the CNN sharing of the target detection. Due to using of the proposed windows CNN and the CNN sharing of target detection, the speed of detection and recognition has been greatly improved. The Faster-RCNN algorithm used in this paper and the algorithm flow chart is shown in Fig. 3.

The CNN network algorithm of Fast-RCNN used in the gesture recognition system is shown in Fig. 4. which owns five layer ZF neural network. The output feature image with the 55*55*96 size is acquired by the the original image data under the processing of 7*7 convolution kernel and 2*2 pool. The step of convolution kernel is 2. Then, the 5*5 and 3*3 convolution kernel are adopted to carry on the further process. Finally, the system outputs a feature image with 13*13*256 characteristics as the input data of the training phase.

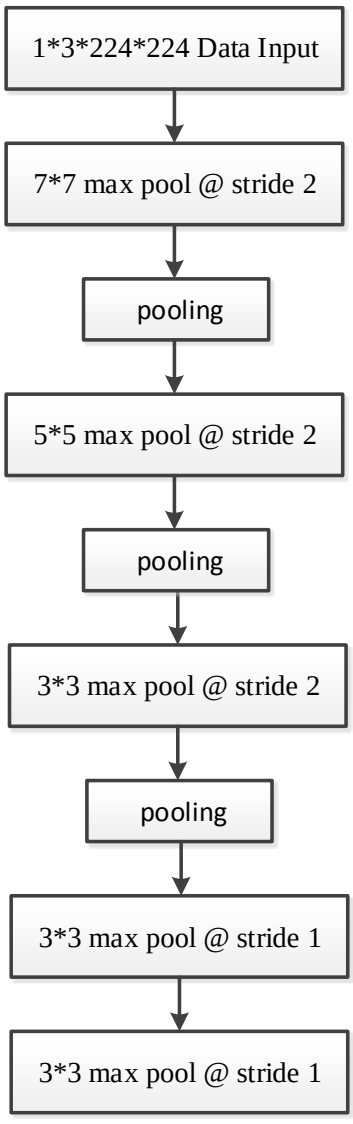

Fig. 4. The flow chart of CNN network.

3. Results and Discussion

The gesture recognition system model was built by using the Faster-RCNN algorithm to train each character gesture with 300 images under three different complex background. The testing result of recognition model character "c", "h", "i", "n" and "a" shows superior performance. It is obvious that that the system has better gesture recognition function. As shown in Fig. 5, the automatic recognition results of character "c" and "h" with deep learning gesture recognition system exhibits high performance. It is known that the two characters can be identified accurately in three backgrounds. what more, the accuracy is better than 90%, and the highest recognition accuracy is up to 99.2%.

The results of the automatic recognition of the characters "i" and "n" with deep learning gesture recognition system processing was observed in Fig. 6. While the results of the automatic recognition of the characters "a" was shown in Fig. 7. It is obvious that that the characters of the "i", "n" and "a" can be accurately recognized under three different complex backgrounds. The recognition accuracy is better than 90% and the highest recognition accuracy is up to 98.8%.

Fig. 5. The automatic recognition result. (a) The recognition result of character “c”; (b) The recognition result of character “h”.

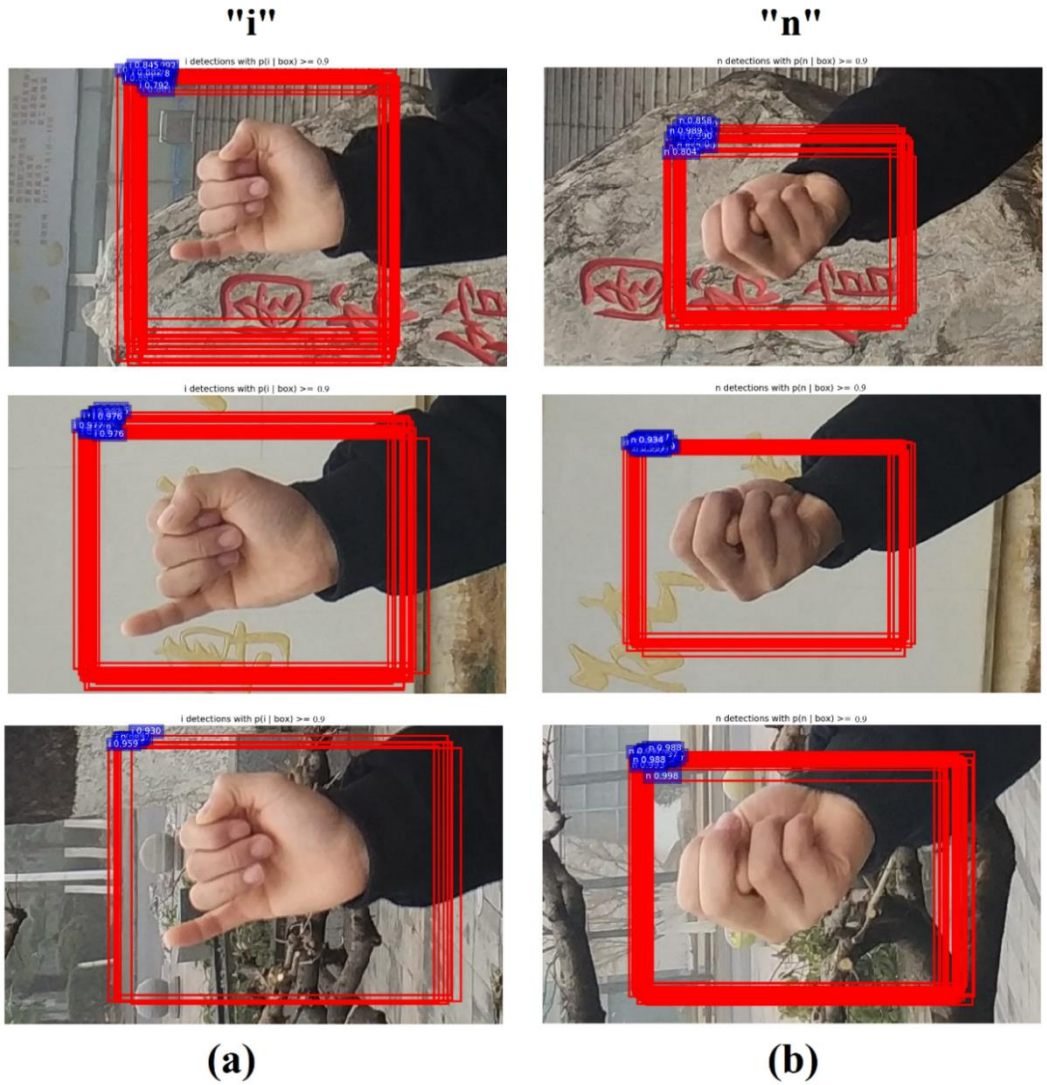

Fig. 6. The automatic recognition result. (a) The recognition result of character “i”; (b) The recognition result of character “n”.

Fig. 7. The automatic recognition result of character “a”.

To evaluate comprehensive performance of the gesture recognition system, the system error and response time were tested. The relationship between the iterations number and the system error was shown in Fig. 8. The system was set on 1000 iterations in the algorithm. It is clearly that the system error decreases with the increasing iteration number. The gesture recognition system average accuracy and response time are shown in Table 1. All the accuracy is better than 90.6% and the "i" characters owns higher accuracy. All the response time is less than 10 ms which is reveal that the system exhibits high real-time performance.

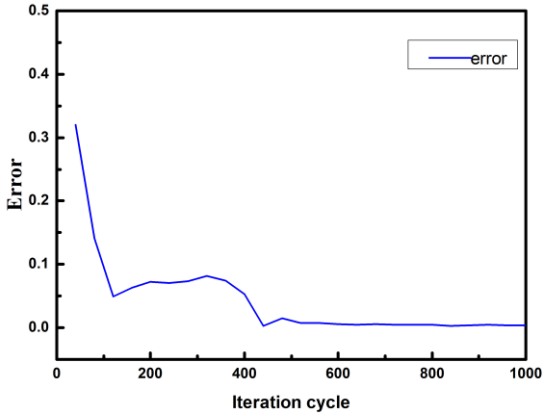

Fig. 8. The relation of iteration number and error.

Table 1. The Accuracy and Response Time of Gesture Recognition System

Character	Accuracy	Response time /ms
"c"	>96.3	5.1
"h"	>96.1	4.7
"i"	>97.5	5.3
"n"	>91.2	7.2
"a"	>90.6	7.5

4. Conclusion

In summary, the faster region full convolution neural network (Faster-RCNN) depth learning algorithm is proposed to apply for gesture recognition in this work. We chose five characters hand gesture under three different complex backgrounds as the investigated objects. The five layers neural network is used to acquire the recognition model with learning and training the selected characters after the hand segmentation carried on. The system test results show that gesture recognition system based on Faster-RCNN exhibits efficiently, reliably, quickly and accurately. The system response time is less than 10 ms revealing high real-time performance. All the accuracy is better than 90.6% and the maximum is high up to 99.2%. The research results show that the Faster-RCNN algorithm can be used in hand gesture recognition for the human-computer interaction application.

Acknowledgement

We acknowledge the financial supports by by National Key R&D Program of China (Grant No. 2017YFE0112000), and Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01).

References

[1] Rautaray, S. S., & Agrawal, A. (2012). Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review, 43(1),* 1-54.

[2] Oyedotun, O. K., & Khashman, A. (2016). Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications, 28(12),* 3941-3951.

[3] Kim, S., Park, G., Yim, S., Choi, S., & Choi, S. (2009). Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction. *IEEE Trans. Consum. Electron, 55(3),* 1169-1177.

[4] Chevtchenko, S. F., Vale, R. F., & Macario, V. (2018). Multi-objective optimization for hand posture recognition. *Expert Systems with Applications, 92,* 170-181.

[5] Mummadi, C., Leo, F., Verma, K., Kasireddy, S., Scholl, P., Kempfle, J., & Laerhoven, K. (2018). Real-time and embedded detection of hand gestures with an IMU-based glove. *Informatics, 5(2),* 28.

[6] Ma, X., & Peng, J. (2018). Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information. *Journal of Sensors,* 5809769.

[7] Xu, D., Wu, X., Chen, Y. L., & Xu, Y. (2014). Online dynamic gesture recognition for human robot interaction. *Journal of Intelligent & Robotic Systems, 77(4),* 583-596.

[8] Hsieh, C. C., & Liou, D. H. (2012). Novel haar features for real-time hand gesture recognition using SVM. *Journal of Real-Time Image Processing, 10(2),* 357-370.

[9] Triesch, J., & Malsburg, C. V. D. (2001). A system for person-independent hand posture recognition against complex backgrounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23(12),* 1449-1453.

[10] Aseema, S., & Rajapuspha, T. (2012). Vision based gesture recognition for alphabetical hand gestures using the SVM classifier. *International Journal of Computer Science & Engineering Technology*, *3(7),* 218-223.

[11] Wang, J., & Wang, G. (2017). Hand-dorsa vein recognition with structure growing guided CNN. *Optik - International Journal for Light and Electron Optics, 149(2017),* 469-477.

[12] Pavlo, M., Shalini, G., Kihwan, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. *IEEE on Computer Vision, 1(3),* 1-7.

[13] Jinn-Tsong, T., Jyh-Horng, C., & Liu, T. K. (2006). Tuning the structure and parameters of a neural network by using hybrid taguchi-genetic algorithm. *IEEE Transaction on Neural Networks, 17(1),* 69-80.

[14] Zhang, C., Tian, Y., Guo, X., & Liu, J. (2018). DAAL: Deep activation-based attribute learning for action recognition in depth videos. *Computer Vision and Image Understanding*, *167(2018),* 37-49.

[15] Barros, P., Parisi, G. I., Weber, C., & Wermter, S. (2017). Emotion-modulated attention improves expression recognition: A deep learning model. *Neurocomputing, 253(2017),* 104-114.

[16] Chen, J., Ou, Q., Chi, Z., & Fu, H. (2016). Smile detection in the wild with deep convolutional neural networks. *Machine Vision and Applications, 28(2),* 173-183.

[17] Shaik, K. B., Ganesan, P., Kalist, V., Sathish, B. S., & Jenitha, J. M. M. (2015). Comparative study of skin color detection and segmentation in HSV and YCbCr color space. *Procedia Computer Science, 57(3),* 41-48.

[18] Tian, Y., Wang, H., & Wang, X. (2017). Object localization via evaluation multi-task learning. *Neurocomputing, 253(2017),* 34-41.

[19] Ren, Y., Zhu, C., & Xiao, S. (2018). Object detection based on fast/faster RCNN employing fully convolutional architectures. *Mathematical Problems in Engineering,* 3598316.

[20] Turan, M., Almalioglu, Y., Araujo, H., Konukoglu, E., & Sitti, M. (2018). Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. *Neurocomputing, 275(2018),* 1861-1870.

[21] Ross, G., Jeff, D., Trevor, D., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(1),* 142-158.

Xiaoguang Yu was born in 1986 Anhui province of China. She received her B.Sc degree in literature art from West Anhui University, China in2006. Then received her M.Sc in digital media system from Wuhan Technology University ,China in 2012. After graduating from Wuhan Technology University she worked as as a lecturer in Department of art at West Anhui University until now. Her research interest include

Image processing, visual art, intelligent control and artificial intelligence.

Yafei Yuan was born in 1984 in China. He received his B.Sc degree in electronic information engineering from Xian Jiaotong University, China in 2006. Then he worked as engineer in Shanghai Institute of Technical Physics of the Chinese Academy of Sciences until 2010. He received his M.Sc in electronic science and technology from China Academy of space technology in 2013. Then he worked as an engineer in China Electronic Technology Corporation (CETC) until 2015. He received his Ph.D in optical engineering from Fudan University in 2018. Currently, he is working in Fudan University as postdoctor. His research interest include signal processing, intelligent optics and artificial intelligence.