Journal of Applied Science and Engineering, Vol. 23, No. 1, pp. 31–38 (2020) *DOI: 10.6180/jase.202003_23(1).0004*

Dynamic Gesture Recognition Based on Deep Learning in Human-to-Computer Interfaces

Jing Yu[1], Hang Li[2]*, Shou-Lin Yin[2]*, Qingwu Shi[4]* and Shahid Karim[3]

[1]*Luxun Academy of Fine Arts, Shenyang 110034, P.R. China*
[2]*Software College, Shenyang Normal University, Shenyang 110034, P.R. China*
[3]*Institute of Image and Information Technology, Harbin Institute of Technology, Harbin 150000, P.R. China*
[4]*College of Information Science & Electronic Technique, Jiamusi University*

Abstract

Currently, gesture recognition provides a faster, simpler, convenient, effective and more natural way for human-computer interaction, which has been widely concerned. Gesture recognition plays an important role in real life. The manual feature extraction in traditional gesture recognition methods is time-consuming and strenuous. Moreover, in order to improve the accuracy of recognition, the quantity and quality of features to be extracted are required to be very high, which is a bottleneck for traditional gesture recognition methods. Therefore, we propose a deep learning method for dynamic gesture recognition in Human-to-Computer interfaces. An improved inverted residual network architecture is utilized as the basis of SSD (Single Shot MultiBox Detector) network for feature extraction. And the convolution structure of the auxiliary layer is predicted by using the inverse residual structure combining the cavity convolution. It uses multi-scale information, which can reduce the amount of calculation and parameters number. Transfer learning is used to optimize the trained network model so as to reduce the training time and make the model more convergent. Finally, experimental results show that the proposed method can recognize different gestures quickly and effectively.

Key Words: Gesture Recognition, Deep Learning, Human-to-Computer Interfaces, Feature Extraction

1. Introduction

The Human-to-Computer interface is a process that people exchange information with computers in a certain way [1]. Recently, gesture interaction based on computer vision has become a research hotspot in the field of Human-to-Computer interfaces due to its convenient and simple equipment. HOG, SIFT and other traditional feature based gesture recognition methods have low recognition accuracy. It is difficult to identify multiple gesture targets in one image [2,3].

Molchanov [4] proposed an algorithm for drivers' hand gesture recognition from challenging depth and intensity data using 3D convolutional neural networks. This method combined information from multiple spatial scales for the final prediction. It also employed spatio-temporal data augmentation for more effective training and to reduce potential overfitting. Wilson [5] extended the standard hidden Markov model method of gesture recognition by including a global parametric variation in the output probabilities of the HMM states. Using a linear model of dependence, it formulated an expectation-maximization (EM) method for training the parametric HMM. Caramiaux [6] presented a gesture recognition/ adaptation system for human--computer interaction applications that, as a complement to gesture labeling, characterized the movement execution. It described a template-based recognition method that simultaneously aligned the input gesture

*Corresponding author. E-mail: lihangsoft@163.com;
352720214@qq.com;
20837445@qq.com

to the templates using a Sequential Monte Carlo inference technique. And many other topics are proposed to detect the gestures. However, there are still some problems such as low efficiency, time-consuming etc.

Deep learning model is a complex, multi-layer artificial neural network structure. Deep learning models have strong nonlinear modeling ability and use general learning process to learn features from data. Compared with the features of traditional artificial design, the deep learning model can express higher level and more abstract internal features [7–9].

Deep convolutional neural network (CNN) in deep learning is an effective method for image feature extraction. Because of its invariance in translation and rotation of image information, it has become a popular method in the field of image processing and target recognition. At present, most of the researches on gesture recognition focus on the gesture recognition with a single hand. In the process of gesture interaction, two-handed operation and other hands often occur. For gesture recognition of multiple hands, this paper proposes a dynamic gesture recognition method based on deep convolutional neural network. Our contributions are as follows:

1. Feature is extracted by an improved inverted residual network architecture based on SSD.
2. The convolution structure of the auxiliary layer is predicted by using the inverse residual structure combining the cavity convolution with multi-scale information, which can reduce the amount of calculation and parameters number.
3. Transfer learning is used to optimize the trained network model so as to reduce the training time and make the model more convergent.
4. Experimental results show that the proposed method can recognize different gestures quickly and effectively.

The rest of this paper is organized as follows. In the next section, we detailed introduce the proposed SSD method for gesture recognition. Then, we give rich experiments and analysis in section 3. A conclusion is conducted in section 4.

2. Gesture Recognition Model in Deep Learning

The main methods of object recognition based on deep convolutional neural network are: RCNN, Fast RCNN, Faster RCNN and SSD, etc. [10–13]. When the PASCAL VOC data set was tested, the object recognition rate of Faster RCNN was 73.2%, and 7 frames of image were recognized in each second. The recognition rate of SSD method was 72.1%, and 58 frames of image were recognized per second. The recognition rate of Faster R-CNN was faster than that of SSD. The recognition rate of YOLO method was 63.4%, and it could recognize 45 frames of image per second. The recognition speed was similar to that of SSD method, and the recognition rate was significantly lower than that of SSD. In this paper, modified SSD (MSSD) model is selected as the recognition model.

2.1 SSD Network Structure

SSD target detection model does not require time-consuming region generation and feature re-sampling steps. By directly convolving the whole image and predicting the category and corresponding coordinates of the object contained in the image, the detection speed is greatly improved. Meanwhile, the accuracy of target detection is greatly improved by using small size convolution kernel and multi-scale prediction.

The SSD network structure is divided into Base network and Auxiliary network. The Base network is the network that has high classification accuracy in the field of image classification and removes its classification layer. The auxiliary network is a convolutional network structure added on the basis of the basic network for target detection. The size of these layers gradually decreases so that multi-scale prediction can be made. Each added auxiliary network layer through a series of convolution kernels will produce a fixed predicted set. For a $m \times n \times p$ (p is the channel number, m, n are the size) feature layer, each auxiliary network will use $3 \times 3 \times p$ convolution kernel to predict and produce score for one class. In the $m \times n$ positions, it predicts all the corresponding values.

SSD model predicts k boundary boxes at each position of feature graph. At the same time, the score of an object appearing in this position and the offset of the object position relative to the boundary box are predicted. Thus, $c \times k$ scores and $4k$ position offsets are predicted at the positions of each feature graph. For a feature graph with $m \times n$ size, it will predict $(c + 4) \cdot k \cdot m \cdot n$ outputs. Finally, non-maximal suppression is applied to obtain the final predicted value of object category and position information in the image.

2.2 Modified SSD Network Structure

SSD model uses VGG network as the basic network. But VGG network model has a large number of parameters, occupies most of the running time in the process of feature extraction. And in the forward propagation process, information loss in the transformation process is always caused by nonlinear transformation.

Shen [14] put forward the nonlinear activation function ReLU based on the manifold learning theory. Under the high dimension, it would be better to retain information. And in the low dimension, it would cause greater loss of information. Therefore, the input layer should increase the feature dimension before the nonlinear transformation. In the output layer, the linear activation function should be used to reduce the dimension of the feature to reduce the loss of information. So inverted residual block was proposed.

The down-sampling operation in the reverse residual structure will cause the loss of feature information while increasing the perceptive field of the convolution kernel. Therefore, it is considered to abandon the down-sampling operation in the convolution structure and introduce the empty convolution to solve this problem. Empty convolution adds an expansion parameter on the basis of the original convolution operation. It expands the convolution kernel to the corresponding scale, and fills 0 in the unused area of the original convolution kernel. The application of empty convolution can increase the sensing field of the convolution kernel without the down-sampling operation. However, the using of empty convolution will make the operation of convolution check data discontinuous, and small objects cannot be better identified. This paper considers the hierarchical feature fusion to solve the problems caused by the introduction of empty convolution.

Hierarchical feature fusion is the sum of the outputs of each convolution unit in the empty convolution layer. And the result of each sum is obtained by concatenate operation to get the final output result.

The reverse residual structure adopts ReLU6 as the activation function, and its output is,

$$Y = \min(\max(X, 0), 6) \tag{1}$$

where Y is the output of ReLU6 activation function. X is the input eigenvalue.

Compared with ReLU, ReLU6 has better robustness in low precision computing scenes. In addition, the convolution kernel of 3×3 is used. Dropout and batch normalization are used in the training network process to reduce the overfitting in the training process. The improved reverse residual structure is shown in Figure 1. Where Dilated denotes the empty convolution, Linear is the Linear activation function, and HFF represents the hierarchical feature fusion. *Dwise* represents a depth-separable convolution structure.

Combining with the improved reverse residual structure, we modify the base layer and auxiliary layer in SSD model: (1) original SSD uses VGG network as a base layer for feature extraction, but VGG network model is not suitable for deployment to run on mobile devices, so reverse residual MobileNetV2 is proposed on the basis of network structure, which has less parameters, small footprint, and running faster, which is as the SSD feature extraction network and to reduce the size of the model and calculation. The traditional convolutional network structure is used in SSD auxiliary layer, which leads to large number of parameters and large amount of calculation. As the basic structure of the auxiliary layer, the improved auxiliary network layer can reduce the information loss caused by the nonlinear transformation in the learning process and the convolution kernel has multi-scale receptive field.

2.3 Loss Function in MSSD Network Structure

Generating recognition box in MSSD model is a re-

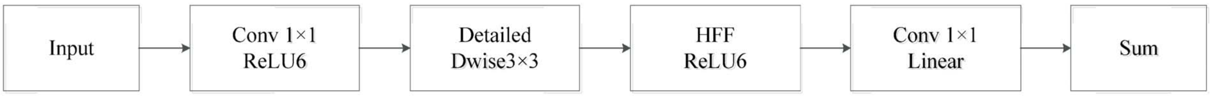

Figure 1. Improved inverted residual network.

gression process. Judging the category within the recognition box is a classification process. The total objective loss function is the weighted sum of position loss (*loc*) and confidence loss (*conf*).

$$L(c, l, g) = 1 / N(L_{conf}(c) + aL_{loc}(l, g)) \tag{2}$$

where, N is the number of default boxes corresponding to the real boxes. $a = 1$ is the weight term according to the real experiment situation. $L_{conf}(c)$ is the cross entropy classification loss function of Softmax, and c is the confidence of each category. In $L_{loc}(l, g)$, $l = (l_x, l_y, l_w, l_h)$, each item denotes the predicted center of the box (x, y) and the width (w), high (h). $g = (g_x, g_y, g_w, g_h)$ represents the true central position (x, y), width (w) and high (h).

$$L_{loc}(l, g) = \sum_{i \in \{x, y, w, h\}} smooth_{L_1}(l_i - g_i) \tag{3}$$

where

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 \ , |x| < 1 \\ |x| - 0.5, \ others \end{cases} \tag{4}$$

3. Experiments on Gesture Recognition

3.1 Data Set Analysis

In order to realize the training of MSSD model, the gesture image data set taken from the first perspective is used. The experiment adopts the gesture data set EgoHands created by Indiana university [15]. The EgoHands use the wearable device Google glass to shoot images. Two people interact with each other in the first perspective. The data set contains 4800 images, and each image contains 4 categories: his own left hand (owlh), his own right hand (owrh), opposite left hand (oplh) and opposite right hand (oprh). Each image labels the gesture region position of 4 categories, as shown in Figure 2.

In training process of MSSD model, the training set, verification set and test set are shown in Table 1.

3.2 Evaluation Index

In this paper, we adopt the following evaluation indexes to analyze the effectiveness of proposed model.

1. IoU (intersection over union) is defined as the ratio of the intersection and union of the area occupied by two boxes [16].

$$IoU = \frac{P \cap GT}{P \cup GT} \tag{5}$$

where P is the predicted box. GT is the ground truth.

2. Precision and recall are two famous quantitative indexes. The gesture recognition model will classify the contents in the identified boxes, predict the possibility of the four gesture categories, and set the most likely as the classification result.

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

$$F - score = \frac{2 Precision \cdot Recall}{Precision + Recall} \tag{8}$$

where TP is the detected correct gesture number. FP is

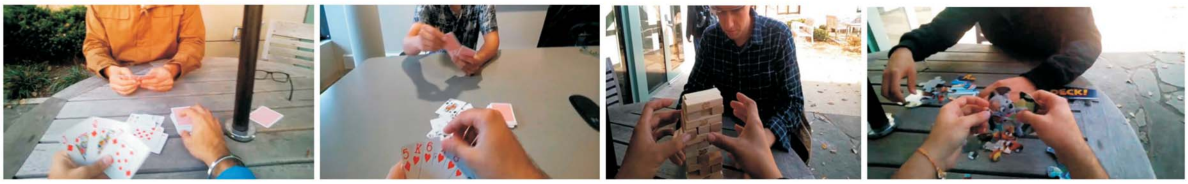

Figure 2. Samples in EgoHands.

Table 1. Some data set in this paper

Data set	Description	Number
Training set	Multiple scenes, multiple people, multiple activities	2500
Verification set	Same with above	500
Test set	Same with above	1000

the detected other posture number. *FN* is the leak detected gesture number. F-score is used to adjust Precision and Recall, which is more close to 1, the model is better.

3. mAP (mean Average Precision) is to get an index that can reflect the global performance.

$$mAP = \int_0^1 Precision(Recall)d\ Recall \qquad (9)$$

3.3 Fine-tuning Network and Transfer Leaning

We firstly verify the effect of IoU on the recognition accuracy with proposed method. The blue bar is accuracy rate of recognition and the red bar is error rate of recognition in Figure 3. When IoU = 0.3, though the recognition rate is high, the error rate is high too. When IoU = 0.6, the result is similar to IoU = 0.9. But IoU = 0.9, it needs more time to process one image. Therefore, we choose IoU = 0.6 in this paper.

For gesture recognition problem in gesture interaction process, the parameters are changed in MSSD model. The VGG-16 recognition model trained in PASCAL VOC dataset is used to initialize the parameters of the basic network in MSSD model. It fixes the first two layers and does not participate in the back propagation. The target to be identified is divided into four categories, and one background category. The total number of categories is set as 5. The maximum recognition results of each frame are set as 4, and the maximum recognition result of each class is set as 1. This set only shows the most likely recognition result in each gesture class, which greatly reduces the false recognition in each class. The training and testing in MSSD model adopt Caffe deep learning framework, and computer graphics card is NVIDIA GTX 1060. The original image size of EgoHands dataset is 1240×720 pixel, which is adjusted to 600×600 during training process. The training strategy is shown in Table 2. In this paper, the fine-tuning and transfer learning are improved in MSSD model network.

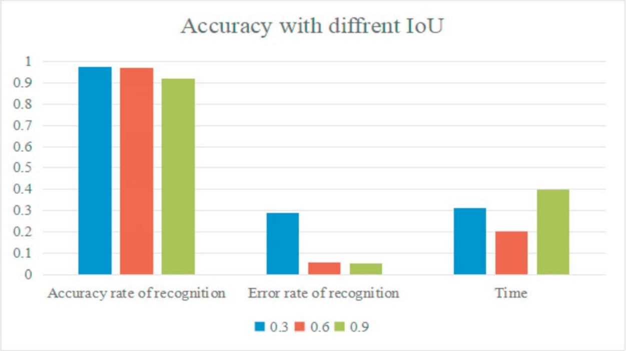

Figure 3. Effect of IoU on recognition.

The size of input image and the size of feature graph with true box would affect the recognition accuracy of MSSD model [17]. The added BN layer will also affect the recognition accuracy of the deep learning model. This experiment will fine-tune the MSSD model structure.

In the experiment, the size of the input image is adjusted from 1240×720 to 600×600 and 300×300. Finally, the trained models are denoted as MSSD6 and MSSD3, respectively. In the experiment, each pixel in the Conv3 ×3 layer extracted from the VGG-16 basic network is added with box. The conv3×3 layer is also introduced into the calculation of loss function and the back propagation process of box recognition, and the training result is MSSD+Conv3 model. The results are shown in Table 3 and Figure 4.

Transfer learning means that a learning algorithm can use the commonalities among different learning tasks to share statistical advantages and transfer knowledge between tasks. Transfer learning can shorten the training time and improve the recognition rate of the model.

Bambach [18] proposed a model for EgoHands gesture recognition based on Caffenet network. In the experiment, the basic network in MSSD model was appropriately changed, and then the parameters in Caffenet model and residual network model (Resnet) were transferred to MSSD model for training.

In the experiment, the MSSD model is adjusted by changing the basic network in VGG as the top-5 layer network in Caffenet model. Then, the parameters of the Caffenet model in [18] are transferred to the basic net-

Table 2. Parameters in SSD model

Name	Value
Size	600 × 600 pixel
Learning rate	10^{-4}
Forgetting rate	0.9
Weight decay	5×10^{-4}
Image number in each iteration	3
Iteration number	64000

Table 3. mAP results with different models

Model	mAP/%	Average recognition image per second
MSSD6	91.3	10
MSSD3	89.5	12
MSSD + Conv3	84.9	9
MSSD + BN	70.8	5

work in the MSSD model to initialize it. Then the network is trained. Training results are as transfer Caffenet model. In addition, the parameters of Caffenet model (top-5 layer structure) are fixed. It dose not participate in reverse back propagation. Training results are as transfer Caffenet top-5 model.

The basic network of MSSD model is changed from VGG to residual network with101 layers. The parameters of the residual network trained in PASCAL VOC data set are transferred to the basic network of MSSD model and initialized. Then the network training is carried out, the training result is as the transfer Resnet101 model. The residual network is relatively complex. In order to shorten the training time, the size of the image is adjusted to 256×256 for training. The training process of each transfer learning model is shown in Figure 5.

Table 4 is the mAP results with different transfer learning methods.

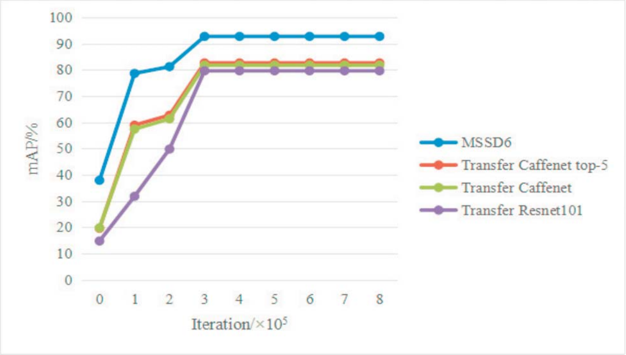

Figure 5. Effect of different transfer learning models on mAP.

Table 4. mAP results with different transfer learning models

Model	mAP/%
MSSD6	92.6
Transfer Caffenet top-5	91.7
Transfer Caffenet	86.2
Transfer Resnet 101	73.4

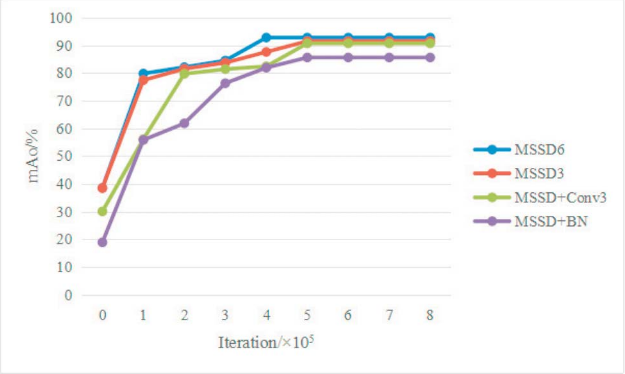

Figure 4. Effect of different models on mAP.

3.4 Comparison experiment

We conduct comparison experiments with other two state-of-the-art dynamic gesture recognition methods including RPS [19], GRM [20], FMCW [21] and LSPD [22]. mAP results are shown in Table 5 and Table 6. Figure 6 displays the mAP value of four hands and Figure 7 presents some gesture recognition results with our proposed method.

Our proposed MSSD method can achieve a better results on all the hands recognition in terms of the mAP. Due to crossed hands with a big area, the recognition re-

Table 5. Comparison results with different methods

Method	Four hands	Precision	Recall	F-score
RPS	owlh	91.73	87.58	89.54
	owrh	92.77	86.31	89.64
	oplh	90.88	85.23	87.46
	oprh	91.25	87.37	89.46
GRM	owlh	92.54	88.71	90.58
	owrh	94.63	90.28	92.86
	oplh	92.86	83.77	88.67
	oprh	93.78	89.65	92.07
FMCW	owlh	93.18	89.67	90.24
	owrh	94.71	90.58	92.45
	oplh	93.14	84.65	87.56
	oprh	94.87	88.56	92.15
LSPD	owlh	94.38	90.84	91.54
	owrh	95.37	91.57	92.84
	oplh	93.94	85.72	89.41
	oprh	95.88	90.63	93.72
MSSD6	owlh	95.21	91.62	93.79
	owrh	96.42	91.08	94.14
	oplh	94.83	90.58	93.18
	oprh	96.88	91.27	94.27

Table 6. mAP results with different methods

Model	mAP/%
RPS	78.6
GRM	84.3
FMCW	84.8
LSPD	85.2
MSSD6	88.9

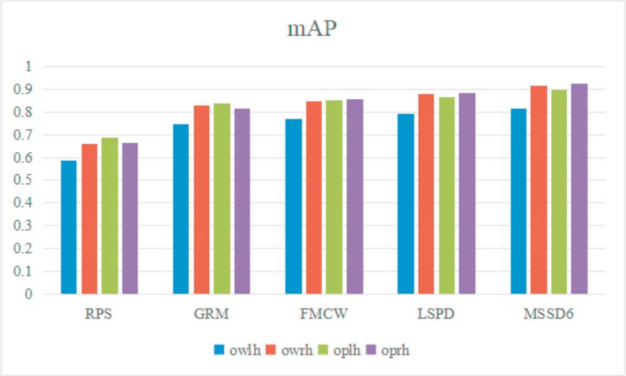

Figure 6. Four hands' mAP value.

sult is not ideal, but it still has productive efficiency than other methods.

Through the wearable device, we take the first view video, and 100 video frames are randomly selected as test images. The mAP obtained on the trained MSSD6 model is 93.2%, and it can recognize 20 pictures per second. The dynamic gesture recognition effect is better.

4. Conclusion

In this paper, we propose a modified deep learning model for dynamic gesture recognition in Human-to-Computer Interfaces. Multiple gestures in the image can be recognized at the same time, the average mAP of gesture recognition with proposed MSSD6 model is larger than 90 percent. It can be used in real-time recognition based on visual gesture interaction. Experiments show that, the method in this paper can quickly and accurately recognize the multi-gesture hands in video. In the future, we will design more advanced CNN network to improve the accuracy of gesture recognition.

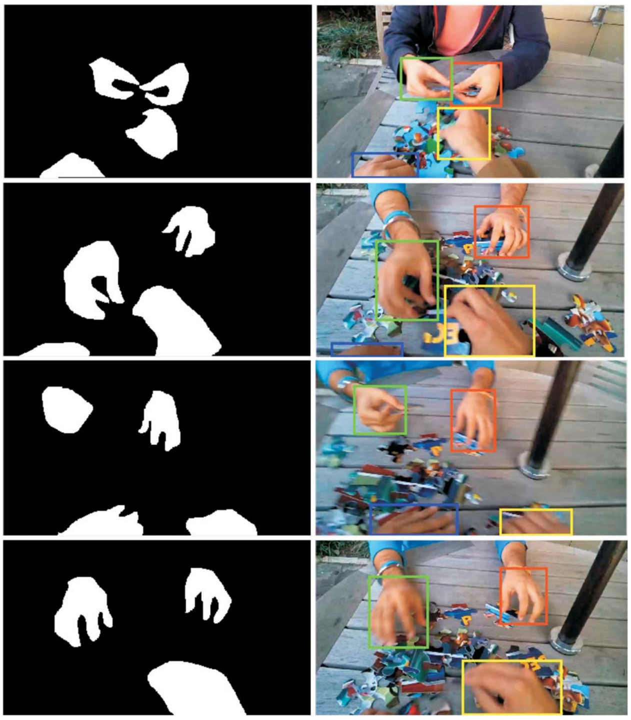

Figure 7. Part of the results: left segment result, right recognition result.

References

[1] Yang, C., J. Long, M. A. Urbin, et al. (2018) Real-time myocontrol of a human–computer interface by paretic muscles after stroke, *IEEE Transactions on Cognitive & Developmental Systems* 10(4), 1126–1132. doi: 10.1109/TCDS.2018.2830388

[2] Mert, A., and A. Akan (2018) Emotion recognition from EEG signals by using multivariate empirical mode decomposition, *Pattern Analysis & Applications* 21(1), 81–89. doi: 10.1007/s10044-016-0567-6

[3] Yu, J., H. Li, and S. L. Yin (2019) New intelligent interface study based on K-means gaze tracking, *International Journal of Computational Science and Engineering* 18(1), 12–20. doi: 10.1504/IJCSE.2019.096971

[4] Molchanov, P., S. Gupta, K. Kim, et al. (2015) Hand gesture recognition with 3D convolutional neural networks, *Computer Vision & Pattern Recognition Workshops*. doi: 10.1109/CVPRW.2015.7301342

[5] Wilson, A. D., and A. F. Bobick (2016) Parametric hidden Markov models for gesture recognition, *IEEE Trans.pattern Anal. & Mach.intell* 21(9), 884–900. doi: 10.1109/34.790429

[6] Caramiaux, B., N. Montecchio, and A. Tanaka (2014) Adaptive gesture recognition with variation estimation

for interactive systems, *Acm Transactions on Interactive Intelligent Systems* 4(4), 1–34. doi: 10.1145/2643204

[7] Gao, J., P. Li, and Z. K. Chen (2019) A canonical polyadic deep convolutional computation model for big data feature learning in Internet of Things, *Future Generation Computer Systems*. doi: 10.1016/j.future.2019.04.048

[8] Lin, T., H. Li, and S. L. Yin (2018) Modified pyramid dual tree direction filter-based image de-noising via curvature scale and non-local mean multi-grade remnant multi-grade remnant filter, *International Journal of Communication Systems* 31(16). doi: 10.1002/dac.3486

[9] Yin, S. L., and J. Bi (2019) Medical image annotation based on deep transfer learning, *Journal of Applied Science and Engineering* 22(2), 385–390. doi: 10.6180/jase.201906_22(2).0020

[10] Yin, S. L., Y. Zhang, and S. Karim (2018) Large scale remote sensing image segmentation based on fuzzy region competition and Gaussian mixture model, *IEEE Access* 6, 26069–26080. doi: 10.1109/ACCESS.2018.2834960

[11] Yin, S. L., Y. Zhang, and S. Karim (2019) Region search based on hybrid CNN in optical remote sensing images under cloud computing environment, *International Journal of Distributed Sensor Networks* 15(5). doi: 10.1177/1550147719852036

[12] Ren, S., K. He, R. Girshick, et al. (2017) Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39(6), 1137–1149. doi: 10.1109/TPAMI.2016.2577031

[13] Li, J., H. C. Wong, S. L. Lo, et al. (2018) Multiple object detection by deformable part-based model and R-CNN, *IEEE Signal Processing Letters* PP(99):1-1. doi: 10.1109/LSP.2017.2789325

[14] Shen, J., J. Bu, B. Ju, et al. (2012) Refining Gaussian mixture model based on enhanced manifold learning, *Neurocomputing* 87(1), 19–25. doi: 10.1016/j.neucom.2012.01.029

[15] Bambach, S., S. Lee, D. J. Crandall, et al. (2015) Lending A hand: detecting hands and recognizing activities in complex egocentric interactions, 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society. doi: 10.1109/ICCV.2015.226

[16] Lepetit-Aimon, G., R. Duval, and F. Cheriet (2018) Large receptive field fully convolutional network for semantic segmentation of retinal vasculature in fundus images, *International Workshop on Computational Pathology* 201–209. doi: 10.1007/978-3-030-00949-6_24

[17] Liu, W., D. Anguelov, D. Erhan, et al. (2016) SSD: single shot MultiBox detector, European Conference on Computer Vision. ECCV, 21–37. doi: 10.1007/978-3-319-46448-0_2

[18] Bambach, S., S. Lee, D. J. Crandall, et al. (2015) Lending A hand: detecting hands and recognizing activities in complex egocentric interactions, 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society. doi: 10.1109/ICCV.2015.226

[19] Zhou, Z., Z. Cao, and Y. Pi (2018) Dynamic gesture recognition with a Terahertz Radar based on range profile sequences and Doppler signatures, *Sensors* 18(1), 10. doi: 10.3390/s18010010

[20] Verma, B., and A. Choudhary (2018) Framework for dynamic hand gesture recognition using Grassmann manifold for intelligent vehicles, *Iet Intelligent Transport Systems* 12(7), 721–729. doi: 10.1049/iet-its.2017.0331

[21] Zhang, Z., Z. Tian, and Z. Mu (2018) Latern: dynamic continuous hand gesture recognition using FMCW radar sensor, *IEEE Sensors Journal* 18(8), 1–1. doi: 10.1109/JSEN.2018.2808688

[22] Nguyen, X. S., L. Brun, O. Lezoray, et al. (2019) Skeleton-based hand gesture recognition by learning SPD matrices with neural networks, IEEE International Conference on Automatic Face & Gesture Recognition (FG). IEEE. doi: 10.1109/FG.2019.8756512

Manuscript Received: Jul. 22, 2019
Accepted: Oct. 19, 2019