Hindawi
Mobile Information Systems
Volume 2019, Article ID 3410348, 7 pages
https://doi.org/10.1155/2019/3410348

Research Article

Hand Gesture Recognition Based on Single-Shot Multibox Detector Deep Learning

Peng Liu,[1] Xiangxiang Li,[2] Haiting Cui,[1] Shanshan Li,[1] and Yafei Yuan[2,3]

[1] *Department of Sports Media and Information Technology, Shandong Sport University, Jinan 250102, Shandong, China*
[2] *Department of Optical Science and Engineering, Fudan University, 200433 Shanghai, China*
[3] *Department of Electronic Engineering, Fudan University, 200433 Shanghai, China*

Correspondence should be addressed to Yafei Yuan; yyf@fudan.edu.cn

Received 20 July 2019; Revised 3 November 2019; Accepted 9 December 2019; Published 30 December 2019

Guest Editor: Malik Jahan Khan

Hand gesture recognition is an intuitive and effective way for humans to interact with a computer due to its high processing speed and recognition accuracy. This paper proposes a novel approach to identify hand gestures in complex scenes by the Single-Shot Multibox Detector (SSD) deep learning algorithm with 19 layers of a neural network. A benchmark database with gestures is used, and general hand gestures in the complex scene are chosen as the processing objects. A real-time hand gesture recognition system based on the SSD algorithm is constructed and tested. The experimental results show that the algorithm quickly identifies humans' hands and accurately distinguishes different types of gestures. Furthermore, the maximum accuracy is 99.2%, which is significantly important for human-computer interaction application.

1. Introduction

With the rapid development of computer technology and artificial intelligence, noncontact gesture recognition plays important roles in human-computer interaction (HCI) applications [1–4]. Due to its natural human-computer interaction characteristics, the hand gesture recognition system allows users to interact intuitively and effectively through a computer interface [5, 6]. Additionally, gesture recognition based on vision is widely applied in artificial intelligence, virtual reality, multimedia, and natural language communication [7–10].

However, traditional hand gesture recognition based on image processing algorithms was not widely applied in HCI because of its poor real-time capacity, low recognition accuracy, and complex algorithm. Recently, gesture recognition based on machine learning has been developed rapidly in HCI due to the introduction of the graphics processor unit (GPU) and the artificial intelligence (AI) image processing [11, 12]. The machine learning algorithms such as local orientation histogram, support vector machine (SVM) [13], neural network, and elastic graph matching are widely used in gesture recognition systems [14–16]. Owning to its learning ability, the neural network does not need manual feature setting during the simulating human learning process and can carry out training the gesture samples to form a network classification recognition map [17, 18]. Deep learning models are inspired by information processing and communication patterns developed from biological nervous systems, which involve neural networks with more than one hidden layer. They can acquire the characteristics of the learning object easily and accurately under the complex object and exhibit superior performance in computer vision (CV) and natural language processing (NLP) [19–21]. Current state-of-the-art object detection systems are variants of Faster R-CNN [22]. The Single-Shot Multibox Detector (SSD) further optimizes object detection [23, 24]. As compared to Faster R-CNN, SSD is more simple and efficient as it completely eliminates proposal generation subsequent pixel and feature resampling stages, and it also encapsulates all computation in a single network which makes SSD easily trainable and straightforward to integrate into systems [5, 25–28].

This paper discusses hand gesture recognition in complex environments based on the Single-Shot Multibox

Detector. The approach is different from the work [28]. The image pyramid method is adapted to gesture recognition. More accurately, the system crops the image into blocks to detect far and small hand gestures. The experiment results show the SSD overcomes the interference signals in complex backgrounds and improves the accuracy and processing speed of gesture recognition.

2. Related Work

Generally, the process of vision-based hand gesture recognition system includes three steps which are hand segmentation, gesture model building, and hand gesture classification. To increase the efficiency, we simplify the process into two steps by using the SSD network. More precisely, we just need a convolutional neural network such as VGG16 [29] as a model system to identify the gesture features and then proceed with hand segmentation and gesture classification simultaneously by the SSD network. This makes our architecture much simpler and much faster than other methods based on the Faster R-CNN model.

The main purpose of gesture model building is to obtain useful semantic features, separate them from the complex backgrounds, and provide effective input information source for the following stage. In the stage of hand segmentation and hand gesture classification, hand postures with different sizes will be located with different bounding boxes. For these bounding boxes, simultaneously, we acquire the confidence for all gesture categories. Training is used for this unified framework 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. In other words, given an input image, we can acquire the location and classification score of hand gesture in this image end-to-end.

The standard hand gesture database is important for the hand gesture recognition system. Figure 1(a) shows the 36 hand gestures from the Massey University's 2D Static hand gesture image dataset which is about standard numbers and letters [30]. Note that some gestures are rather difficult to distinguish from each other. For example, "a" and "e," "d" and "l," "m" and "n," or "i" and "j." In this paper, we have chosen the characters of "w," "o," "r," and "k" as the study objects which are shown in Figure 1(b). The Canon EOS 6D camera was employed to capture the gesture with an EF 24–105mm/4L IS USM lens and a shutter time of 1/100 S. And the maximum distance is about five meters. Each hand gesture sample was obtained under three different complex backgrounds, aiming to prove the applicability and reliability of the hand gesture recognition system.

The hand gesture model building plays a vital role in a gesture recognition system that is regarded as the first step for processing the original input gestures. The inputs of this stage are images. When seeing an image, from the perspective of human beings, we can catch the sight of the scene described in the picture. However, the computer cannot capture these scenes from an original picture. The computer thinks an image is just a matrix with a variety of values in different spatial locations and channels. In other words, the computer can only obtain pixel-level information of an image. Obviously, it is difficult to distinguish different objects using low-level information such as pixel values. Therefore, if we want to recognize hand gestures, one of the most efficient methods is extracting and summarizing high-level information such as their features and structures from the original image. This is exactly what gesture modeling does in our framework. We use the VGG16 convolutional neural network, which uses 13 convolutional layers and is deep enough to obtain high-level information of hand gestures. Given the original image as the input, the VGG16-Net will output feature maps of different resolutions which contain high-level information of the image. The reason for choosing 19 layers is that it is enough to extract high-level semantic information for classification and regression. And limited by the size of our dataset, using high-level layers can easily lead to overfitting.

The VGG-Nets are a series of convolutional neural networks with different depths which all use very smaller (3 × 3) convolution filters. The VGG16-Net (16 weight layers) is one of them which has 13 convolutional layers and 3 fully connected layers. The structure of VGG16 is shown in Figure 2. In this figure, the convolutional layer parameters are denoted as "conv < receptive field size > − < number of channels>." The ReLU activation function is not shown for brevity. The original image is passed through a stack of convolutional layers, which use filters with a small receptive field: 3 × 3 (which is the smallest size for capturing the notion of the left, right, up, down, and center). The convolution stride is fixed to 1 pixel; the spatial padding of a convolutional layer is such that the spatial resolution is preserved after convolution, i.e., the padding is 1 for 3 × 3 convolution filters. Spatial pooling is carried out by five maxpooling layers, which follow some of the convolutional layers (not all the convolutional layers are followed by maxpooling layers). Maxpooling is performed over a 2 × 2 pixel window, with stride 2.

All convolutional layers are equipped with the rectification nonlinearity (ReLU) [31]. After a stack of convolutional, maxpooling, and ReLU layers, we get feature maps with lower resolution and stronger semantic information. There are also fully connected layers and a soft max layer which are used for image classification in the original VGG16-Net. We replace these layers with SSD layers to implement hand segmentation and hand gesture classification.

The second stage, i.e., using the SSD network to perform hand segmentation and hand gesture classification, is the most important part in our framework. We have chosen the SSD model because it is both accurate and fast. The core of SSD is predicting category scores and bounding box offsets for a fixed set of default bounding boxes using very small (3 × 3) convolutional filters applied to feature maps. Beyond that SSD produces predictions of different scales from feature maps of different scales and separates predictions by aspect ratio. This architecture leads to simple end-to-end training and high accuracy, further improving the speed versus accuracy trade-off [5].

SSD is based on a feed-forward convolutional neural network (VGG16) that produces a fixed-size collection of bounding boxes and scores for the presence of object class

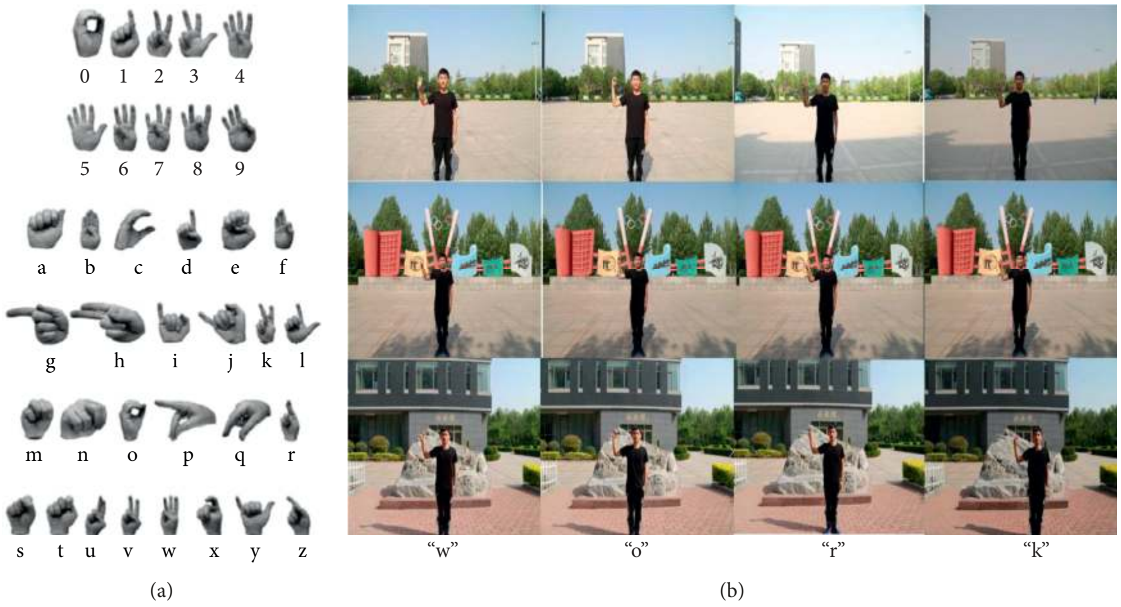


Figure 1: The graphs of hand gesture information: (a) standard library of hand gestures and (b) four hand gestures graphs in different backgrounds.

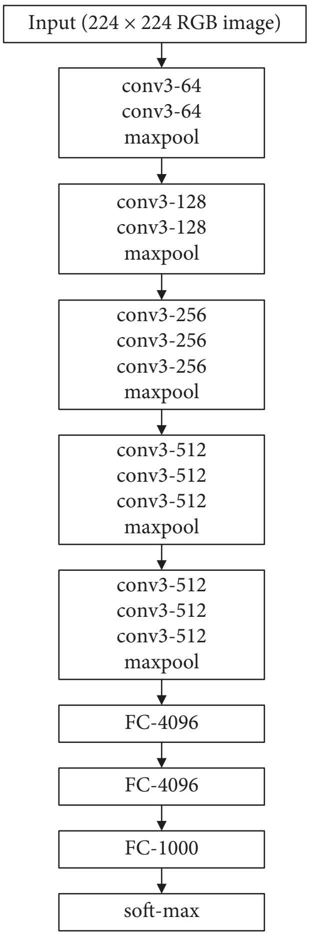


Figure 2: VGG16 Conv-Net structure.

instance in those boxes. This approach will produce a large number of bounding boxes, and most of them are covered by each other. Therefore, a nonmaximum suppression step is executed to discard repetitive bounding boxes and produce the final detections. The structure of SSD is shown in Figure 3. The input image is an image with 300 × 300 pixels

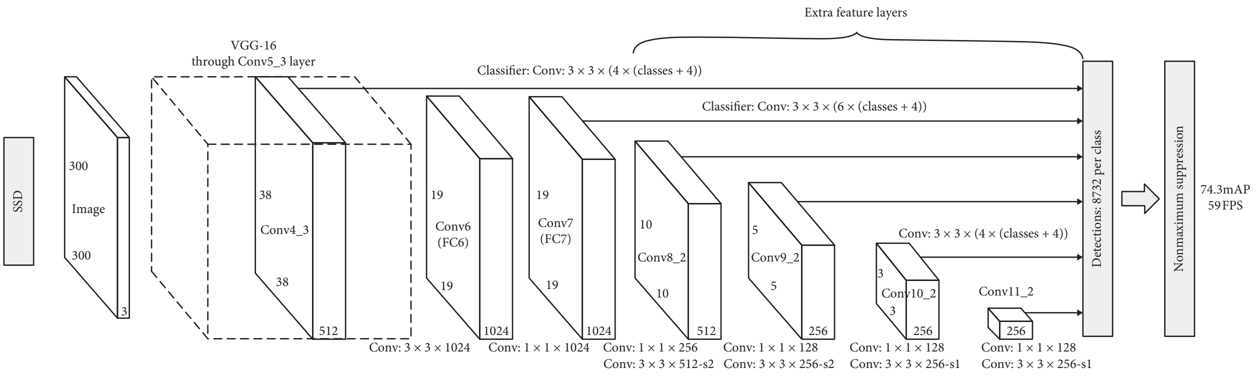


Figure 3: Structure of the SSD network.

and RGB channels. The part in the dotted box is the truncated VGG16 network. The SSD model adds several feature layers of different scales to the truncated VGG16 network. These layers decrease in size as depth increases and allow predictions of detections at multiple scales. Then, small convolutional filters apply to every position in selected feature maps. More precisely, these filters apply to a set of default boxes of different aspect ratios at each location in several selected feature maps to predict the shape offsets and the confident scores for all object categories. In our work, object categories include four hand gestures and the background.

Noting that we have the SSD framework, the next thing we need is an objective function to train the model end-to-end. The overall objective function is a weighted sum of the localization loss (loc) and the confidence loss (conf):

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

where N is the number of default boxes that match to ground truth boxes. The localization loss is a smooth $L1$ loss between the ground truth box (g) and the predicted box (l) parameters. These parameters are offsets for the center coordinate (cx, cy) of the default bounding box (d) and for its width (w) and height (h), which is similar to Faster R-CNN [22]:

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

The confidence loss is the soft max loss over multiple class confidences (c), as is usually used in multiple classification tasks:

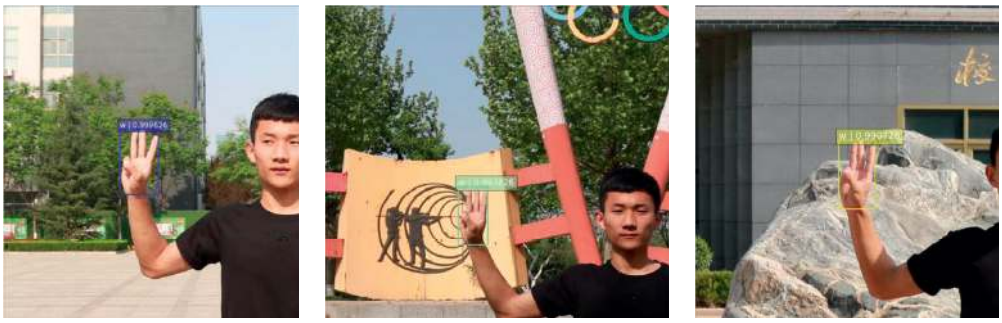
$$\begin{aligned} L_{\text{conf}}(x, c) = &- \sum_{i \in \text{Pos}}^{N} x_{ij}^{p} \log\left(\hat{c}_i^p \right) \\ &- \sum_{i \in \text{Neg}} \log\left(\hat{c}_i^0 \right), \text{ where } \hat{c}_i^p = \frac{\exp\left(c_i^p \right)}{\sum_p \exp\left(c_i^p \right)}. \end{aligned} \tag{3}$$

During training, we match the default boxes to the ground truth boxes to calculate and reduce the loss of objective function. We do this recursively to optimize the parameters of the SSD model and finally get an ideal model. By using k-means clustering to guide the aspect ratio of anchor boxes, we get three different ratios. After that the ratios are 1.9, 1.6, and 1.1 with slight adjustment, respectively. Furthermore, the used optimizer is Adam with an initial learning rate of 0.0001.

3. Results and Discussion

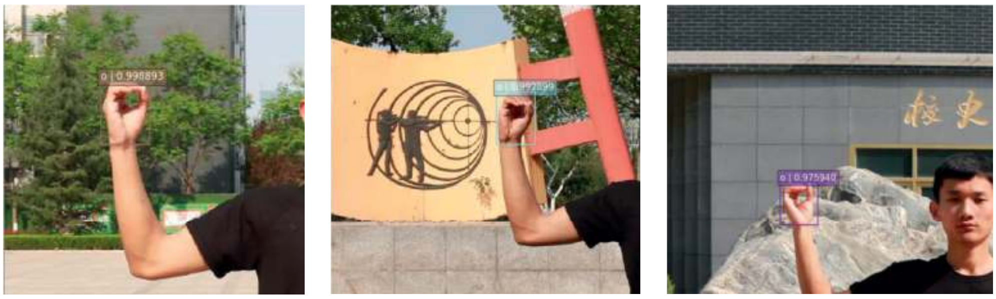
The hand gesture recognition system was built by the SSD algorithm and training each character gesture with 1070 images with three different complex backgrounds. Then, we used 268 images which were not in the training set to test the building recognition model. The testing results of the recognition model on characters “w,” “o,” “r,” and “k” show good performance. In all 268 images, 261 of them are recognized correctly, with an accuracy of more than 93.8% and the highest recognition accuracy of 99.2%. The average prediction confidence for the 261 images recognized successfully is up to 0.96, which is very close to 1. Examples of visualization results are shown in Figures 4–7 with the character “w,” “o,” “r,” and “k,” respectively.

To evaluate the comprehensive performance of the gesture recognition system, the recognition accuracy for each hand gesture and response time was tested. The average accuracy of the gesture recognition system and response time are shown in Table 1. All the accuracies are more than 93.8%, and the character “o” owns higher accuracy. All



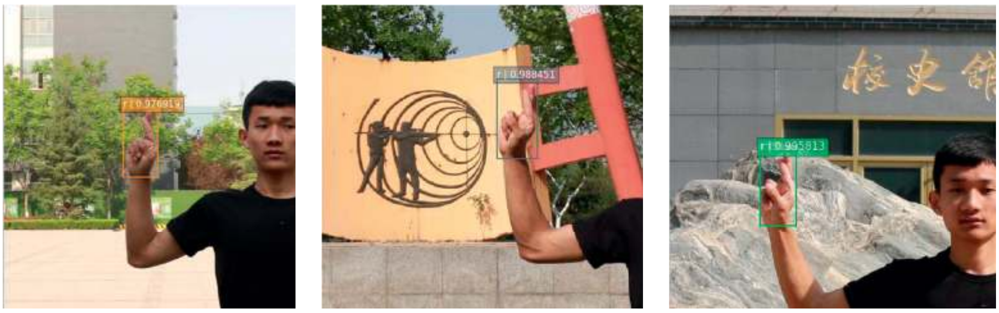
(a) (b) (c)

Figure 4: The automatic recognition result of character "w."



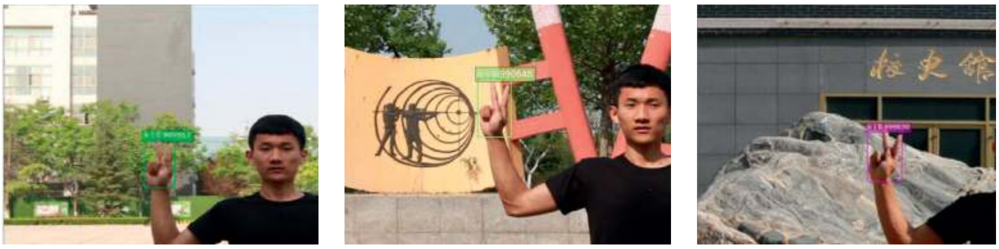
(a) (b) (c)

Figure 5: The automatic recognition result of character "o."



(a) (b) (c)

Figure 6: The automatic recognition result of character "r."



(a) (b) (c)

Figure 7: The automatic recognition result of character "k."

Table 1: The accuracy and response time of the gesture recognition system.

Character	Accuracy	Response time (ms)
"w"	>96.9	17.1
"o"	>99.1	18.3
"r"	>93.8	18.2
"k"	>98.6	19.7

Table 2: Recognition accuracy comparison for the English alphabets.

Works	Method	Training number	Test number	Accuracy (%)
Liu et al. [32]	Baum–Welch and Viterbi Path Counting	520	260	85.77
Sahoo et al. [33]	Firefly-based back propagation	442	442	73.3
Kundu et al. [34]	26-point feature extraction and ANN	520	260	86.5
Zeng et al. [35]	Leap motion via deterministic	2340	2600	92.9

response times are less than 20 ms which shows that the system exhibits high real-time performance.

The proposed work contributes to promote the accuracy of the hand gestures recognition as alphabets ("w," "o," "r," and "k") with the employment of SSD and image cropping. The results show that the adopted classification approach exhibits superior performance, which clearly indicates that the proposed system is an effective method for the hand gestures recognition. It is found, by comparing with other works, that the accuracy of the proposed method adopted in our work is higher than that of others which are listed in Table 2.

4. Conclusion

The Single-Shot Multibox Detector (SSD) deep algorithm is proposed to apply to the hand gesture recognition. We chose four character's hand gestures under three different complex backgrounds as the investigated objects. The 19-layer convolutional neural network is used as a recognition model with learning and training the selected characters end-to-end. The system test results show that the hand gesture recognition system based on the SSD model performs efficiently, reliably, quickly, and accurately. The response time of the system is less than 20 ms revealing high real-time performance. The minimum accuracy is more than 93.8%, and the maximum is 99.2%. The research results show that the SSD algorithm can be used in the hand gesture recognition system for the human-computer interaction application.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the financial supports by the Shandong Key Research and Development Plan Project (grant no. 2019GGX105018), National Key R&D Program of China (grant no. 2017YFE0112000), and Shanghai Municipal Science and Technology Major Project (grant no. 2017SHZDZX01).

References

[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2012.

[2] S. Kim, G. Park, S. Yim, S. Choi, and S. Choi, "Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1169–1177, 2009.

[3] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2016.

[4] J. Jeong and Y. Jang, "Max–min hand cropping method for robust hand region extraction in the image-based hand gesture recognition," *Soft Computing*, vol. 19, no. 4, pp. 815–818, 2014.

[5] W. Liu, D. Anguelov, D. Erhan, S. Szegedy, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," *Computer Vision—ECCV 2016*, vol. 21, pp. 21–37, 2016.

[6] Y. Wu, D. Jiang, X. Liu, R. Bayford, and A. Demosthenous, "A human-machine interface using electrical impedance tomography for hand prosthesis control," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 6, pp. 1322–1333, 2018.

[7] X. Ma and J. Peng, "Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information," *Journal of Sensors*, vol. 2018, Article ID 5809769, 9 pages, 2018.

[8] S. F. Chevtchenko, R. F. Vale, and V. Macario, "Multi-objective optimization for hand posture recognition," *Expert Systems with Applications*, vol. 92, pp. 170–181, 2018.

[9] C. Mummadi, F. Leo, K. Verma et al., "Real-time and embedded detection of hand gestures with an IMU-based glove," *Informatics*, vol. 5, no. 2, p. 28, 2018.

[10] D. Xu, X. Wu, Y.-L. Chen, and Y. Xu, "Online dynamic gesture recognition for human robot interaction," *Journal of Intelligent & Robotic Systems*, vol. 77, no. 3-4, pp. 583–596, 2014.

[11] J. Dongarra, M. Gates, J. Kurzak, P. Luszczek, and Y. M. Tsai, "Autotuning numerical dense linear algebra for batched

computation with GPU hardware accelerators," *Proceedings of the IEEE*, vol. 106, no. 11, pp. 2040–2055, 2018.

[12] M. Morchid, "Parsimonious memory unit for recurrent neural networks with application to natural language processing," *Neurocomputing*, vol. 314, pp. 48–64, 2018.

[13] C.-C. Hsieh and D.-H. Liou, "Novel Haar features for real-time hand gesture recognition using SVM," *Journal of Real-Time Image Processing*, vol. 10, no. 2, pp. 357–370, 2012.

[14] A. Sultana and T. Rajapuspha, "Vision based gesture recognition for alphabetical hand gestures using the SVM classifier," *International Journal of Computer Science & Engineering Technology*, vol. 3, no. 7, pp. 218–223, 2012.

[15] J. Triesch and C. von der Malsburg, "A system for person-independent hand posture recognition against complex backgrounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1449–1453, 2001.

[16] J. Wang and G. Wang, "Hand-dorsa vein recognition with structure growing guided CNN," *Optik*, vol. 149, pp. 469–477, 2017.

[17] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," *IEEE on Computer Vision*, vol. 1, no. 3, pp. 1–7, 2015.

[18] J.-T. Tsai, J.-H. Chou, and T.-K. Liu, "Tuning the structure and parameters of a neural network by using hybrid Taguchi-genetic algorithm," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 69–80, 2006.

[19] J. Chen, Q. Ou, Z. Chi, and H. Fu, "Smile detection in the wild with deep convolutional neural networks," *Machine Vision and Applications*, vol. 28, no. 1-2, pp. 173–183, 2016.

[20] C. Zhang, Y. Tian, X. Guo, and J. Liu, "DAAL: deep activation-based attribute learning for action recognition in depth videos," *Computer Vision and Image Understanding*, vol. 167, pp. 37–49, 2018.

[21] P. Barros, G. I. Parisi, C. Weber, and S. Wermter, "Emotion-modulated attention improves expression recognition: a deep learning model," *Neurocomputing*, vol. 253, pp. 104–114, 2017.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[23] J. Yu, J. Li, B. Sun, J. Chen, and C. Li, "Multiclass radio frequency interference detection and suppression for SAR based on the single shot multibox detector," *Sensors*, vol. 18, no. 11, p. 4034, 2018.

[24] A. Fuentes, S. Yoon, S. Kim, and D. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," *Sensors*, vol. 17, no. 9, p. 2022, 2017.

[25] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a surface defect detection algorithm based on MobileNet-SSD," *Applied Sciences*, vol. 8, no. 9, p. 1678, 2018.

[26] Y. Wang, C. Wang, and H. Zhang, "Combining a single shot multibox detector with transfer learning for ship detection using sentinel-1 SAR images," *Remote Sensing Letters*, vol. 9, no. 8, pp. 780–788, 2018.

[27] X. Zhao, W. Li, Y. Zhang, and Z. Feng, "Residual super-resolution single shot network for low-resolution object detection," *IEEE Access*, vol. 6, pp. 47780–47793, 2018.

[28] C. Yi, L. Zhou, Z. Wang, Z. Sun, and C. Tan, "Long-range hand gesture recognition with joint SSD network," in *Proceedings of the IEEE Internal Conference on Robotics and Biomimetics*, Kuala Lumpur, Malaysia, December 2018.

[29] D. Zhao, D. Zhu, J. Lu, Y. Luo, and G. Zhang, "Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16," *Symmetry*, vol. 10, no. 10, p. 519, 2018.

[30] A. L. C. Barczak, N. H. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2d static hand gesture colour image dataset for ASL gestures," *Research Letters in the Information and Mathematical Sciences*, vol. 15, pp. 12–20, 2011.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106–1114, 2012.

[32] N. Liu, B. C. Lovell, P. J. Kootsookos, and R. I. A. Davis, "Model structure selection & training algorithms for an HMM gesture recognition system," in *Proceedings of the IEEE 9th International Workshop on Frontiers in Handwriting Recognition*, pp. 100–105, Tokyo, Japan, October 2004.

[33] M. K. Sahoo, J. Nayak, S. Mohapatra, B. K. Nayak, and H. S. Behera, "Character recognition using fire-fly based back propagation neural network," *Computational Intelligence*, vol. 2, pp. 37–49, 2016.

[34] S. Kundu, H. S. Chhabra, S. S. Ara, and R. P. Mishra, "Optical character recognition using 26-point feature extraction and ANN," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 5, pp. 156–162, 2017.

[35] W. Zeng, C. Wang, and Q. Wang, "Hand gesture recognition using leap motion via deterministic learning," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28185–28206, 2018.

Applied Computational Intelligence and Soft Computing
The Scientific World Journal
Mathematical Problems in Engineering
Journal of Engineering
Advances in Multimedia
Modelling & Simulation in Engineering
Advances in Artificial Intelligence
International Journal of Reconfigurable Computing
Advances in Fuzzy Systems
Scientific Programming
Advances in Human-Computer Interaction
International Journal of Engineering Mathematics
Advances in Civil Engineering
Journal of Computer Networks and Communications
Journal of Robotics
International Journal of Computer Games Technology
International Journal of Biomedical Imaging
Journal of Electrical and Computer Engineering
Computational Intelligence and Neuroscience