Research Status of Gesture Recognition Based on Vision: A Review

Jun Tian¹, *, Weilie Zhang¹, Tao Zhang¹, Guopeng Li¹ and Zhang Long²
¹School of Engineering, Southwest Petroleum University, Nanchong, China
²Nanchong Key Laboratory of Robotics Engineering and Intelligent Manufacturing, Southwest Petroleum University, Nanchong, China
*Corresponding author e-mail: juntian@swpu.edu.cn

Abstract. Gesture control, as a new type of interactive method, has the characteristics of rich expression, convenient control, and fast. It has huge application prospects in entertainment, home furnishing, and industry. Gesture recognition is the basis of gesture control. Gesture recognition technology based on visual detection acquires gesture information in a non-contact manner, which enables the operator to have a better operating experience and is favored by scholars at home and abroad. In order to fully understand the existing research methods of visual gesture recognition, firstly, the basic process of visual gesture recognition is explained. According to the principle of the gesture recognition method, it is divided into gesture recognition based on traditional methods and gesture recognition based on deep learning. And the specific methods are analyzed and summarized in detail. Finally, the technical difficulties of visual gesture recognition are analyzed and discussed, and the development trend of gesture recognition based on vision is prospected.

1. Introduction

Human-computer interaction technology is gradually changing from computer-centered to human-centered. As an important human-computer interaction method, gesture control provides people with a natural and intuitive communication mode. It has been initially applied in the entertainment industry and has strong practicality in robot control [1]. Gesture recognition is the basis of gesture control. At present, gesture recognition is mainly divided into gesture recognition based on the data glove as the input device and gesture recognition based on the visual device as the input device [2, 3]. Based on visual gesture recognition technology to obtain gesture information in a non-contact manner [4], so that the operator has a better operating experience, it is favored by scholars at home and abroad. In addition, the data glove device is expensive, making the gesture control method of the data glove gradually eliminated in recent years.

Early visual gesture recognition relied on smearing or pasting easy-to-recognize color or shape marks on the human hand, and realized gesture recognition through visual recognition marks [5]. This method is difficult to recognize rich gestures due to the limited quantity and quality of marks. With the development of visual technology, unmarked visual gesture recognition has become the current mainstream. The general sense of vision-based gesture recognition is unmarked visual gesture
recognition. The following description is unmarked visual gesture recognition unless otherwise specified.

The main process of visual gesture recognition includes: (1) Image acquisition: use a visual camera to collect gesture images; (2) Hand detection and segmentation: detect the position of the hand in the gesture image and segment the hand region; (3) Gesture recognition: extraction the image features of the hand area, and the gesture type is recognized based on the features. The process of visual gesture recognition is shown in Figure 1.

![Figure 1. The process of vision-based gesture recognition.](image)

Vision-based gesture recognition technology can be divided into static gesture recognition and dynamic gesture recognition [6]. Static gesture recognition takes the hand in a static state for recognition, and the posture, shape, position and other data information of the hand will not change [7]. It has the advantage of high recognition efficiency. However, static gesture recognition also has its own shortcomings. For example, static gestures can only express less information, which does not match the characteristics of actual human hand movement. Dynamic gestures are composed of static gestures frame by frame, that is, static gestures are a special state of dynamic gestures [8]. The advantage of dynamic gesture recognition is that constantly changing gestures can express more information and can be widely used in the field of human-computer interaction.

According to the principle and development process of gesture recognition, it can be divided into gesture recognition based on traditional methods and gesture recognition based on deep learning. This paper studies and analyzes the literature related to visual gesture recognition in recent years, and outlines the common methods of gesture recognition. Then, the main technologies involved in each process of gesture recognition are compared and analyzed. These processes include gesture detection and segmentation, gesture tracking, gesture feature extraction, and gesture classification. Finally, this article also discusses the challenges and limitations of gesture recognition, as well as possible future development directions to promote research in this field.

2. Gesture recognition based on traditional methods

Researchers of gesture recognition based on traditional methods mainly rely on traditional machine learning methods and image processing methods for gesture recognition related work. Most gesture recognition methods include two stages, namely detection and recognition.

2.1. Gesture detection

The first task of gesture recognition is the detection of hand location and the segmentation of the image area. In traditional methods, hand segmentation is crucial because it isolates task-related data from the complex image background, and then the segmented data is passed on to the subsequent recognition stage. References [9, 10] proposed several types of visual features and their combination methods in most cases. These characteristics include the skin color, shape, movement, and anatomical model of the hand.

The method of gesture detection based on skin color mainly selects a suitable color space (such as RGB, HSV, etc.) to establish a skin color model [11]. Because skin color-based detection will be interfered by illumination changes, in order to enhance the invariance of illumination changes, some
methods operate on color spaces such as HSV [12] and YUV [13]. Generally speaking, the method of hand region detection and segmentation by color has great limitations, and it is susceptible to interference from similar objects in the background.

The detection method based on the shape is mainly by capturing the shape characteristics of the hand in the image [14]. A large number of methods use edge detection to extract the contour of the hand. In general, the contour extraction method based on edge detection will also detect a large number of edges belonging to other objects, which requires a complicated post-processing process. Therefore, edge extraction is often combined with skin color and background separation. The local topological descriptor is used to match the model with the edges in the image. Argyros et al. [15] established the corresponding relationship based on the three-dimensional position information of multiple fingertips, obtained the three-dimensional information of the hand, and reconstructed the three-dimensional contour of the detected and tracked hand in real time.

2.2. Gesture recognition

The overall goal of gesture recognition is to interpret the position, posture, or semantics conveyed by the gesture. In order to detect static gestures, a classifier or template matcher is generally used. However, dynamic gestures have a time dimension, and it is necessary to use methods to deal with this dimension, such as Hidden Markov Model (HMM) [16], to model the time dimension through gesture representation.

In the recognition based on static gestures, methods such as random forest, support vector machine (SVM), neural network (NN), etc. are mainly used. Huang et al. [17] combined Gabor filter and SVM to classify 11 static gestures. Pugeaultp et al. [18] combined the hand shape features extracted from the RGB image with the depth image, and used the random forest method on the 24 gesture data set ASL (American Sign Language) to achieve a recognition rate of 75%.

In the recognition based on dynamic gestures, a large number of studies [19, 20] use hidden markov model (HMM) for recognition. The dynamic gesture is defined as a set of state sequences. In the process of gesture recognition, each state can represent a set of possible hand positions or postures. State transition means that a certain hand position or posture is transferred to another position or posture. The corresponding output symbol indicates a specific gesture.

3. Gesture recognition based on deep learning

Traditional machine learning methods have great limitations when dealing with unprocessed data [21]. Machine learning researchers need considerable professional domain knowledge, perform complex preprocessing of the task data that needs to be performed, design a corresponding feature extractor, and further convert the original data image information into feature vectors, and then input the resulting feature vectors The corresponding classifier to output the target category. Then the obtained feature vector is input into the corresponding classifier to output the target category. Deep learning can gradually express the original data as high-level features by combining simple and nonlinear modules [22]. In this way, deep learning can learn very complex feature representations.

At the same time, with the development of deep learning technology in recent years and the gradual maturity of deep convolutional neural network (CNN), especially the proposed convolutional network models [23, 24] such as VGG Net, Google Net, Res Net, Dense Net, etc., significant breakthroughs have been made in image classification and recognition tasks. Moreover, due to the powerful learning ability of deep learning when dealing with end-to-end problems, it can handle more complex tasks, such as various target detection, video intelligent monitoring, human body pose estimation, etc. In recent years, there has been a lot of work using deep learning methods for gesture recognition, which is also the current mainstream method.

3.1. Gesture recognition based on 2D-CNN

Two-dimensional convolutional neural networks (2D-CNN) are mostly used to process static gestures or process dynamic gesture sequences frame by frame, and its network structure is shown as in Figure 2. Kang et al. [25] used CNN to extract features from the fully connected layer for gesture recognition
of depth images. Oyedotun and Khashman et al. [26] used CNN and multi-layer noise reduction autoencoder (SDAE) to recognize 24 static gestures. Liang et al. [27] proposed a multi-view framework to recognize gestures from point cloud data, by projecting the point cloud model of the hand to different view planes, and then using CNN to extract features from these views. Koller et al. [28] extracted millions of weakly labeled data from sign language videos, and used Google Net's Inception-V1 model to train 60 gesture classifiers, then combined the CNN classifier with HMM to recognize gestures in the video sequence.

3.2. Gesture recognition based on 3D-CNN

Some three-dimensional convolutional neural network (3D-CNN) models have been proposed for gesture recognition. The common network structure of 3D-CNN model is shown in Figure 3. Molchanov et al. [29, 30] proposed a 3D-CNN model to recognize car driving gestures based on depth and intensity data, and combined multiple spatial scale information to make the final prediction, and used a loop mechanism to expand the 3D-CNN model to detect and classify dynamic gestures. Its network model includes 3D-CNN to extract spatio-temporal features and a loop layer for global timing modeling. Reference [31] improved the 3D-CNN model proposed by Tran et al. [32], using depth and RGB information for large-scale gesture recognition. Similarly, Camgoz et al. [33] also built an end-to-end 3D-CNN model based on the Tran model for large-scale gesture recognition. Miao et al. [34] combined 3D-CNN with Res Net and proposed the ResC3D model, which extracted features from time series RGB images, optical flow data, and depth data, and then classified them through SVM, and achieved 67.71% recognition rate on the Chalerm LAP IsoGD dataset.

![Figure 2. 2D-CNN gesture recognition network framework diagram.](image)

![Figure 3. 3D-CNN gesture recognition network framework diagram.](image)
The advantage of 3D-CNN is that it can dynamically recognize gestures, using the fusion momentum of normalized depth and image gradient values, and increase spatiotemporal data to avoid over-fitting. The combination of low-resolution and high-resolution sub-networks significantly improves classification accuracy. The disadvantage is that gestures are limited to specific scenarios, insufficient information processing for serialization, and insufficient robustness to complex environments.

3.3. Gesture recognition based on RNN-LSTM
Neverora et al. [35] first applied recurrent neural network (RNN) to gesture recognition, and its network structure is shown as in Figure 4. For depth, skeleton pose, and speech, a multi-modal gesture recognition system is proposed. Each mode is processed separately in a short time sequence. The features of each mode are manually extracted or learned, and then a long-term dependence model is established through RNN to use for data fusion and final classification. Eleni et al. [36] proposed convolutional long short-term memory recurrent neural network model (RNN-LSTM), which can successfully learn gesture features of different duration and complexity.

![Figure 4. RNN-LSTM gesture recognition network framework diagram.](image)

The advantage of using RNN-LSTM for gesture recognition is that it proposes a method of duration classification, which can continuously recognize dynamic gestures, and has a good processing process for sequence information such as video, making the classification of gestures more accurate and real-time. However, its disadvantage is that the construction model and pre-training are complicated, and it is still not robust to the recognition of complex environments.

4. Technical difficulties and development trends of gesture recognition

4.1. Technical difficulties of gesture recognition
Gesture recognition technology has developed rapidly in recent years. However, due to the interference of external environmental factors and various limitations of the gesture itself, it is easy to have various effects on the system, making gesture recognition still have insurmountable problems. In order to make the way of human-computer interaction using gesture recognition more perfect, this paper summarizes the following technical difficulties, which are used as reference for related scientific research workers.
(1) Gesture segmentation in complex background
   At present, most methods have achieved recognition in isolated scenes, but the complex background environment factors are changeable. When the illumination changes, the background has similar skin colors, etc., the detection, tracking and segmentation of gestures in the video will bring great difficulties.

(2) Gesture recognition for occlusion or blind spots
   When a human hand moves in space, there are often problems of objective occlusion or gesture recognition under blind spots. This situation brings great difficulties to gesture tracking and gesture recognition. BALLAN L et al. proposed to use differentiated learning on the fingers, and associate the distinguishing point features of the fingers with gestures [37], achieving extremely accurate recognition in the case of interaction between hands and hands, and hands and objects. However, due to the large number of conditions required for identification and the large amount of calculation of the model, it is not applicable in actual human-computer interaction applications.

(3) Variety of gestures
   The human hand has 27 degrees of freedom, which is equivalent to a highly flexible deformable body. The movement of the hand includes changes in position and rotation. Therefore, the human hand can make a variety of complex and different gestures, which greatly increases the difficulty of hand feature analysis.

(4) Real-time gesture recognition
   In the process of recognizing the gesture model, since the input is a video image sequence, when the image pixels are high and the amount of processed data is large, the system needs to be able to process a large amount of data quickly, which has high requirements for computer hardware Requirements. This has also become a difficult problem for real-time gesture recognition.

4.2. The development trend of gesture recognition
   With the expansion of the application range of gesture recognition, the requirements for the richness and real-time performance of gestures are getting higher and higher, making dynamic gesture recognition one of the research hotspots. And complex gesture recognition is another research direction of gesture recognition.

   With the rise of artificial intelligence, deep learning is undoubtedly a benign accelerator for gesture recognition. If gesture recognition based on deep learning can be fully integrated into the human-computer interaction system, then human-computer interaction will become more natural and efficient, while can also improve the intelligence of human-computer interaction to a large extent, this will definitely be the future development trend.

5. Conclusion
   Vision-based gesture recognition has an irreplaceable role in the field of human-computer interaction. After years of extensive research, vision-based gesture recognition has made great progress, but because gestures are usually in a complex environment, and complex background will cause the accuracy of gesture segmentation to decrease, it cannot be accurately extracted and recognized. Gesture. This paper investigates the main methods of vision-based gesture recognition in recent years, reviews several representative methods of gesture recognition based on traditional methods and gesture recognition based on deep learning, summarizes the advantages and disadvantages of various methods, and points out the current research technical difficulties and development trends. With the development of human-computer interaction and the maturity of research in deep learning and other fields, reliable, accurate, and effective gesture recognition systems will surely be developed to improve people's lives.

Acknowledgments
   This work was financially supported by the Nanchong city and school science and technology strategic cooperation project (Grant No.19SXHZ0041, No.19SXHZ0036 and No.19SXHZ0037) fund.
References

[1] K Qian, Niu, H Yany. Developing a gesture based remote human-robot interaction system using kinect [J]. International Journal of Smart Home, 2013, 7 (4): 203-208.

[2] S Bhowmick, A K Talukdar, K K Sarma. Continuous hand gesture recognition for English alphabets [C] // International Conference on Signal Processing and Integrated Networks, IEEE, 2015: 443-446.

[3] D Tang, H J Chang, A Tejani, et al. Latent regression forest: structured estimation of 3d hand poses [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2017, 39 (7): 1374-1234.

[4] S Q Ren, K M He, R Girshick, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137-1149.

[5] R Y Wang, J Popovic. Real-time hand-tracking with a color glove [J]. ACM Transactions on Graphics, 2009, 28 (3): 1-8.

[6] A Rautara, A Agrawal. Vision based hand gesture recognition for human computer interaction: a survey [J], Artificial Intelligence Rev, 2015 (43): 1-54.

[7] Y Ji, S Kim, K B Lee. Sign language learning system with image sampling and convolutional neural network [C] // 2017 First IEEE International Conference on Robotic Computing (IRC), Taiwan: IEEE, 2017: 371-375.

[8] M Elbadawy, A S Elons, Shedeed H A, et al. Arabic sign language recognition with 3D convolutional neural networks [C] // 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo: IEEE, 2017: 66-71.

[9] M Cote, P Payeur, G Comeau. Comparative study of adaptive segmentation techniques for gesture analysis in unconstrained environments [C] // IEEE international workshop on imaging systems and techniques. 2006: 28-33.

[10] X Zabulis, H Baltzakis, A Argyros. Vision based Hand gesture recognition for human-computer interaction [C] // The Universal Access Handbook LEA, 2009.

[11] R A Elsayed, M S Sayed, M I Abdalla. Skin-based adaptive background subtraction for hand gesture segmentation [C] // 2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS). Cairo, Egypt: IEEE, 2015: 33-36.

[12] D Saxe, R Foulds. Toward robust skin identification in video images [C]. Proceedings of the Second International Conference on. Automatic Face and Gesture Recognition, IEEE, 1996: 379-384.

[13] J Yang, W Lu, A Waibel. Skin-color modeling and adaptation [G], Asian Conference on Computer Vision. Springer, Berlin, Heidelberg, 1998: 687-694.

[14] B Yan, Y Li, S T Ren, et al. Recognition and evaluation of corrosion profile via pulse-modulation eddy current inspection in conjunction with improved Canny algorithm [J]. NDT & E International, 2019, 106: 18-28.

[15] A Argyros, M I A Lourakis. Binocular hand tracking and reconstruction based on 2D shape matching [C]. Pattern Recognition, 2006. ICPR2006.18th International Conference on. IEEE, 2006:1: 207-210.

[16] K Oka, Y Sato, H Koike. Real-time fingertip tracking and gesture recognition [J]. Computer Graphics and Application IEEE, 2002, 22 (6): 64-71.

[17] D Y Huang, W C Hu, S H Chang. Vision-based hand gesture recognition using PCA +Gabor filters and SVM [C]. Fifth International Conference on Intelligent Information Hiding & Multimedia Signal Processing. IEEE Computer Society, 2009: 1-4.

[18] N Pugeault, R Bowden. Spelling it out: Real-time ASL fingerspelling recognition [C]. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE Computer Society, 2011: 1114-1119.

[19] F S Chen, C M Fu, C L Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models [J]. Image and Vision Computing, 2003, 21 (8): 745-758.
[20] M Malgireddy, I Nwogu, V Govindaraju. Language -motivated approaches to action recognition [J]. The Journal of Machine Learning Research, 2013, 14 (1): 2189-2212.

[21] T S Furey, N Cristianini, N Duffy, et al. Support vector machine classification and validation of cancer tissuesamples using microarray expression data[J]. Bioinformatics, 2000, 16 (10): 906.

[22] P Bao, A I Maqueda, C R Del-Blanco, et al. Tiny hand gesture recognition without localization via a deep convolutional network [J]. IEEE Transactions on Consumer Electronics, 2017, 63 (3): 251-257.

[23] A Krizhevsky, I Sutskever, G E Hinton. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60 (6): 84-90.

[24] H Qassim, A Verma, D Feinzimer. Compressed residual-VGG16 CNN model for big data places image recognition [C]// IEEE Computing & Communication Workshop & Conference. IEEE, 2018, 169-175.

[25] B Kang, S Tripathi, T Q Nguyen. Real -time sign language finger spelling recognition using convolutional neural networks from depth map [C]. Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on. IEEE, 2015: 136-140.

[26] O K Oyedotun, A Khashman. Deep learning in vision -based static hand gesture recognition [J]. NeurD Computing and Applied, 2017, 28 (12): 3941-3951.

[27] C Liang, Y Song, Y Zhang. Hand gesture recognition using view projection from point cloud [C]. Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016: 4413-4417.

[28] O Koller, H Ney, R Bowden. Deep hand: how to train a CNN on 1 million hand images when your data is continuous and weakly labelled [C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3793-3802.

[29] P Molchanov, S Gupta, K Kim, et al. Hand gesture recognition with 3D convolutional neural networks [C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 1-7.

[30] P Molchanov, X Yang, S Gupta, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4207-4215.

[31] Y Li, W Li, V Mahadevan, et al. Vlad3: Encoding dynamics of deep features for action recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1951-1960.

[32] D Tran, L Bourdev, R Fergus, et al. Learning spatiotemporal features with 3d Convolutional networks [C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[33] N C Camgoz, S Hadfield, O Koller, et al. Using convolutional 3d neural net-works for user-independent continuous gesture recognition [C]. Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016:49-54.

[34] Q Miao, Y Li, W Ouyang, et al. Multimodal gesture recognition based on the ResC3D network [C]. IEEE International Conference on Computer Vision Workshop (ICCVW). 2017: 3047-3055.

[35] N Neverova, C Wolf, G Paci, et al. A multiscale approach to gesture detection and recognition [C]. Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013: 484-491.

[36] E Tsironi, P Barros, S Wermer. Gesture recognition with a convolutional long short-term memory recurrent neural network [J], Bruges, Belgium, 2016, 2: 15-23.

[37] L Ballan, A Tanja, L V Gool, et al. Motion capture of hands in action using discriminative salient points [C]// European Conference on Computer Vision. S. l. Springer-Verlag, 2012: 640-653.