Research on Gesture Recognition Method Based on Computer Vision

Xianghan Wang[1,a], Jie Jiang[1,b*], Yingmei Wei[1,c], Lai Kang[1,d], Yingying Gao[1,e]

[1] Department of Systems Engineering, National University of Defense Technology, Changsha 410000, China;

Abstract. Gesture recognition is an important way of human-computer interaction. With time going on, people are no longer satisfied with gesture recognition based on wearable devices, but hope to perform gesture recognition in a more natural way. Computer vision-based gesture recognition can transfer human feelings and instructions to computers conveniently and efficiently, and improve the efficiency of human-computer interaction significantly. The gesture recognition based on computer vision is mainly based on hidden Markov, dynamic time rounding algorithm and neural network algorithm. The process is roughly divided into three steps: image collection, hand segmentation, gesture recognition and classification. This paper reviews the computer vision-based gesture recognition methods in the past 20 years, analyses the research status at home and abroad, summarizes its current development, the advantages and disadvantages of different gesture recognition methods, and looks forward to the development trend of gesture recognition technology in the next stage.

1 Introduction

Gesture recognition is an important part of computer science with the aim of understanding human gestures through algorithms. Computer vision-based gesture recognition enables people to communicate more naturally with machines. The advantage is that it is less affected by the environment. Users can interaction with computer at any time, and it has less constraint on users, enabling computers to accurately and timely understand user's instruction. The instructions do not require any mechanical assistance. Gestures are timely, vivid, intuitive, flexible and visual in the process of human-computer interaction. They can soundlessly complete interaction and successfully break the gap between reality and virtual.

Early gesture recognition directly detected the position of each joint of the human hand with wearable devices, and transmitted the information to the computer through wired transmission, accurately storing the user's hand motion information. For example, Data Gloves and other equipment, although the detection effect is very good, but they are expensive and inconvenient to use. Subsequently, the optical marking method uses infrared light to detect the position and hand movement of the human hand, and it displaced Data Glove. It also has a good effect, but still requires more complicated equipment. Although higher accuracy can be obtained by the help of external devices, it is expensive and affects the user's actions to some extent. The gesture recognition based on computer vision refers to the processing of the video data collected by the camera through the gesture recognition algorithm, which achieves the purpose of gesture recognition and has become a research hotspot in recent years.

2 Research status at home and abroad

At present, gesture recognition based on computer vision has become a research hotspot in the field of computer vision. Grobel and Assan[1] used HMM (Hidden Markov Model) to identify isolated gestures in the video, with a recognition accuracy of 94% for 262 gestures. Reyes and Dominguez[2] proposed a method of gesture recognition using DTW (Dynamic Time Warping), which completes the recognition of depth gesture images in video. Simo-Serra et al.[3] added physical constraints to the joint points of hand to finish gesture recognition. Ayan et al.[4] proposed a method using the Matrix Completion[5] in 2016, which can be applied to large-scale real-time gesture pose estimation without relying on GPU.

Domestic computer vision-based gesture recognition has also achieved certain achievements. Zhu Yuanxin of Tsinghua University[6] proposed a method of gesture recognition using apparent changes in image transformation, and used variational parameter model of image motion to identify 120 gestures. Xiao Ling[7] of Hunan University proposed a method to solve dynamic gesture recognition through self-learning sparse representation. This method directly processes the original image without feature extraction and real-time processing. Liu Shuping[8] of the University of Science and Technology of China proposed a method for finger

[b*]Corresponding author: jiejiang@nudt.edu.cn

detection of static gestures. The feature extraction was performed by HOG to obtain a binary image. The number of fingers was obtained by logical open operation, and finally classified by SVM (Support Vector Machine). The method combines the method of binary image and gray image, and the number of fingers can be reduced to narrow the possible range of the gesture to be recognized. The accuracy of 25 gestures on the data set is about 99%, but the data set is relatively simple and not utilized, didn't use complex gestures for verification. Wang Kai et al[9] proposed a real-time gesture recognition scheme combining optical flow matching and AdaBoost algorithm, which only needs to read 2D video clips to get accurate results.

At home and abroad, a lot of researches has been done on gesture recognition. Different models were used to solve gesture problems. Studies on static, dynamic, isolated and continuous gestures have been done. But there are still problems of poor versatility and intense dependence on data sets.

3 Key Technology of Gesture Recognition Based on Computer Vision

Gesture recognition based on the computer vision obtains the picture containing the human hand by cameras, and then utilizes the image processing and the machine learning to contribute to the judgment and recognition of the gesture comprehensively. At present, the steps of gesture recognition based on computer graphics are split into three stages: image collection, hand detection and segmentation, gesture recognition and classification.

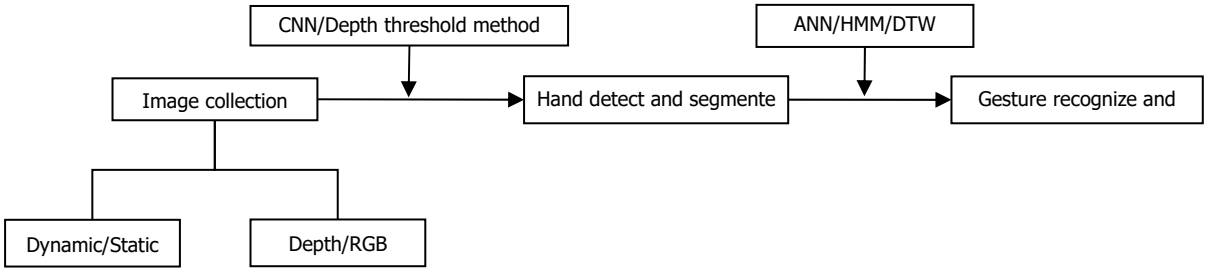

Figure 1. The process of gesture recognition

3.1 Hand detection and segmentation

Image collection is divided into depth images and RGB images. RGB images can be got with a normal camera, and depth cameras such as Kinect and Leap Motion have the ability to simultaneously acquire depth images and RGB images. The use of depth images enables the recording of part of the spatial information, which facilitates the recognition and classification of gestures. Gestures are divided into static gestures and dynamic gestures, which determine whether get a single photo or a video.

In the process of image collection, there are problems such as occlusion, different light intensity and direction, which put forward higher requirements for the robustness of the algorithm. With the put forward of practicality of gesture recognition, more and more algorithms are devoted to achieving illumination invariance and dealing with occlusion problems.

3.1.1 Gesture Segmentation Based on Convolutional Neural Networks

Gesture segmentation based on convolutional neural networks includes optimization through convolutional neural networks, based on FCN (Full Convolutional Neural Networks) [10] or SegNet [11]. FCN and SegNet replace the last layer of CNN with a deconvolution layer, and the image is restored to its original size by up sampling, and each pixel is predicted. Compared to CNN, FCN and SegNet can accept input images of any size, no longer require all images to have the same size, and avoids the problem of repeated storage and convolution calculations. However, FCN also has obvious shortcomings. When the up sampling factor is high, the image is not very clear, and the sensitivity to detail needs to be improved; the relationship between different pixels is not fully utilized. Markus[12] added a convolutional neural network to predict the middle finger joint of the palm in the hand region extraction method, which significantly improved the division effect. Christian[13] uses SegNet to segment hand area. According to the divided area, take a screenshot of the area near the hand, the robustness of the division is improved, but the obtained image contains the edge area, which affects the recognition accuracy, as shown in picture 2. Wei et al. [14] combined the full convolutional neural network with context semantics to improve the accuracy of segmentation.

The segmentation method based on convolutional neural networks is flexible, and a variety of different methods can be combined to perform gesture segmentation.

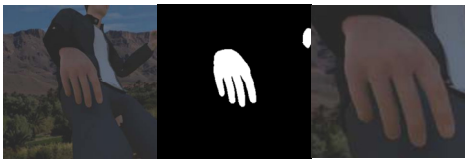

Figure 2. Hand detection and segmentation[13]

3.1.2 Gesture Segmentation Based on Depth Threshold Method

The depth threshold method gets the distance between each pixel and the camera according to the distance between the object and the camera in the depth image, and extracts an image whose distance is within a

predetermined range. In order to better extract the hand range, the depth range is defined for the hand on the depth image, or the hand is directly considered to be the object closest to the camera. This method improves the pre-processing effect, obtains a more accurate hand area, and improves the accuracy of gesture recognition. However, it limits the recognition range and has limitations in the recognition process. Ayan [4] uses the method of maximum spot detection, selects an image with a depth of [50,500] mm, and adds a wristband to the hand to distinguish between the hand and the background. The combined use of multiple methods guarantees the accuracy of the divided gestures, but imposes certain restrictions on the user's actions. Markus [15] considers the hand as the closest object to the depth sensor, uses the depth threshold to determine the hand area, and uses the hand center of mass as the center of the extracted hand area. Aiming at the problem that the interference of noise in the gesture image makes the boundary of the gesture unclear, Li Qing et al.[16] proposed a method of gesture segmentation using the improved maximum inter-class method. The method locates noise by the gray histogram of the gesture image, effectively improve segmentation accuracy. Mao Yanming et al. [17] added conditional constraints after segmentation by depth threshold method to reduce the impact of wrists and fingers drawing close. It is an effectively solution to reduce the impact of wrist and finger draw close on gesture recognition.

Although the depth threshold method can perform image segmentation simply and efficiently, it has a large constraint on user behaviour, and the improvement space is not very large.

3.2 Gesture Recognition and Classification

Gestures recognition are divided into dynamic recognition and static recognition. Static gestures are gestures of a single picture, while dynamic gestures are changes in gesture movements over time, that is, multiple consecutive static gestures. The image of gesture recognition is divided into depth map, RGB map and RGB-D map. The depth map can directly represent the distance between the camera and the object. The depth map's representation is similar to the gray image. The difference is that the depth map shows the distance between the object and the camera with every pixel. The RGB-D image contains RGB three-channel RGB images and depth images. Although the two images look different, there is a one-to-one correspondence between the pixels.

In recent years, gesture recognition is mostly based on artificial neural networks (ANN-CNN, RNN, GAN), HMM and DTW algorithms. The DTW and HMM algorithms were originally used in speech recognition. The DTW algorithm is based on the DP (dynamic programming) idea to solve the problem of different lengths of pronunciation. DTW algorithm does not require a large amount of data to train like HMM and the convolutional neural network. The algorithm is simple and fast, the idea is to find the optimal path and find the matching sequence with the smallest overhead according to the optimal path. HMM looks for hidden sequences from observable sequences and finds the meaning to be expressed by gestures. The convolutional neural network was originally used for image classification. In gesture recognition, it can be used for feature extraction and dimensionality reduction, or directly classification. It is more flexible but requires a large amount of labelled training data. It takes time in the training process and needs GPU acceleration.

3.2.1 HMM-based Gesture Recognition

Zhang et al. [18] proposed an improved HMM dynamic gesture recognition algorithm based on B-parameters, optimized the B-parameter training process, which made it doesn't fall into local optimum easily, simplified the training process, and enhanced the illumination invariance. Wang Xiying et al.[19] of the Institute of Software, Chinese Academy of Sciences proposed a model of fusion of HMM and FNN, which can make full use of human experience to help modelling and classification, and can effectively complete the recognition of complex dynamic gestures. The disadvantage is that the stereo gesture is decomposed into three directions, resulting in no data integrity. Zhang et al. [20] proposed an HMM gesture recognition method based on template matching, which enhances the illumination invariance and background interference robustness in the process of gesture segmentation based on depth image. Xu Jiabin[21] proposed a gesture recognition based on the parallel HMM method. The Fourier transform is used to mine the time domain and frequency domain features, and the isolated gestures are split, which simplifies the calculation. Yu Meijuan et al.[22] proposed a method to combine HMM with dynamic programming, which reduces the computational complexity of HMM method and improves the accuracy and real-time of interaction. Wang Wanliang et al.[24] proposed a gesture recognition method combining HMM with wearable devices. The advantage is that it reduces the limitation of external factors such as occlusion and light, and can realize real-time recognition, but is constrained by wearable devices.

HMM has been widely used as a typical method of probability and statistics. It was first used in the field of speech recognition. In recent years, it has also developed greatly in gesture recognition. HMM-based gesture recognition requires the establishment of an HMM for each gesture, which is computationally intensive and affects real-time performance.

3.2.2 Neural Network-Based Gesture Recognition

In recent years, neural network-based methods have become mainstream. Franziska et al.[24] proposed a method for generating 3D human hand models from RGB pictures in real time using the improved GAN (Generative Adversarial Nets)[25]. Gestures are generated by GAN, and 2D coordinates and 3D coordinates are obtained through multi-objective learning. Spurr et al. [26]

proposed a method for gesture recognition in potential space through VAE frame and cross-modal KL divergence. It can perform semi-supervised learning and can perform gesture recognition on both RGB images and depth images. In 2017, Christian, Thomas et al.[13] proposed a method for 3D estimation using one RGB image of the hand. They used PoseNet[27], which calibrates the camera position according to the scene, the hand position is calibrated according to the camera position, and the joint position of the hand is interpreted in combination with the prior knowledge to accurately acquire the bone node through single RGB image. The disadvantage is that the entire network takes a long time, and there is still a lot of optimization space, so that it is hoped to achieve real-time recognition on the CPU. In 2015, Markus et al.[15] proposed a method based on convolutional neural network for preliminary positioning and re-optimization. It can accurately locate the hand on a single depth image after training with multiple labelled depth images. The method of node location is named HandDeep. On the basis of HandDeep, Markus et al.[12] proposed DeepPrior++ in 2017. Based on the original model, DeepPrior++ increases the total sample size by stretching and scaling the image. The original convolutional neural network is optimized to ResNet deep neural network to improve the recognition accuracy. A new gesture segmentation is proposed. The method, the segmentation is more accurate, and the constraint on the position of the gesture is reduced.

The algorithm based on neural network is flexible and can be adjusted in many aspects. It can be used in combination with different algorithms. However, the general network structure is extremely complex, and the parallel computing capability of the GPU is relatively high.

3.2.3 DTW-based Gesture Recognition

Chen Junjie et al[28] proposed a weighted DTW gesture recognition algorithm. Through the study of the correlation of gestures, the weighting of the hand nodes is completed, and the classification ability of the DTW cost function is significantly improved. The result is better than DTW and HMM. Yu Chao et al.[29] of the University of Science and Technology of China proposed a dynamic gesture recognition technology based on TLD (Tracking-Learning-Detecting) and DTW, which effectively solved the problem of poor stability of gesture tracking and low recognition efficiency. Li et al.[30] proposed a method for measuring the similarity of gesture motion trajectory by DTW algorithm, performing template matching and realizing fast and accurate gesture recognition. He Chao et al[31] proposed a comprehensive DTW algorithm and weighted distance and global distance method to enhance the illumination invariance and real-time. Zhou Zhiping[32] proposed a method of dynamic gesture authentication combining DTW and STLCS (Short Time Longest Common Subsequence).The method based on DTW needs less computation. But the accuracy is lower than that of neural network.

4 Problems and Trends

After more than 20 years of experience, the computer vision-based gesture recognition technology has made great progress. It has gone through the optical flow method, SVM classification method, adding constraint method, HMM algorithm, DTW algorithm and ANN algorithm. However, it still faces many constraints, and there are still certain bottlenecks in development. There is a problem:

(1) Seriously affected by the background. Similar to traditional image classification, good gesture segmentation is still required in complex backgrounds. When gesture recognition is performed, whether the gesture can be accurately separated from the background is the key to improving the recognition accuracy;

(2) Occlusion problems. In dynamic gesture recognition, gestures may be blocked by some objects in the environment, increasing the difficulty of gesture tracking;

(3) Different gestures are similar, and the same gesture may be different;

(4) Multiple degrees of freedom, hand has a lot of freedom. There are many degrees of freedom in the hand activity space. It is difficult for various algorithms to perform satisfied calculation for each degree of freedom, and it takes a long time to calculate multiple degrees of freedom, which increases the difficulty of real-time recognition.

(5) Different viewing angles and light intensity. Rotation invariance and illumination invariance are more difficult to perform in the gesture recognition process.

Compared with many methods, the neural network method is slow, accurate, and dependent on data. It requires a large amount of tagged data and powerful computing speed to ensure that it cannot meet real-time requirements. Compared with the HMM method, the DTW method is faster than the HMM method, but the accuracy and model robustness are not as good as those of the neural network. As shown in Table 1.

Table 1. Comparison of different methods about gesture recognition

Algorithm	Number of samples required	Calculate speed	Recognition accuracy	Number of papers recent years
ANN	Extremely large	Extremely slow (usually need GPU for speed up)	High	Many
DTW	Small	Faster	Low	A few
HMM	Large	Middle	Middle	A few

So far, neural network-based methods have become mainstream algorithms, especially those based on convolutional neural networks. The recognition accuracy is high, the robustness is good, and it can adapt to dynamic and static; depth, RGB image and other different conditions. Gesture recognition requirements. Not only can gesture segmentation, but also gesture recognition and feature extraction, and dimensionality reduction. The gesture recognition method based on the cyclic neural network and the generated confrontation network is also adopted by more and more people as an emerging gesture recognition algorithm. This paper believes that the improvement of computer operation speed can solve the problem of poor real-time neural network, and the application of neural network will become more and more widely. Using SegNet or FCN for gesture segmentation, and then using CNN for dimensionality reduction and feature extraction, and finally using RNN (Recurrent Neural Network) or GAN to complete the identification will become the mainstream solution.

5 Conclusion

This paper reviews the gesture recognition based on computer vision. After nearly 20 years of development, gesture recognition has broken through the wearable constraints. But there are still problems such as poor universality, sensitivity to illumination changes and occlusion, and poor real-time performance. Indeed, the improvement of computer operation speed can solve the problem of poor real-time performance. The influence of poor universality and illumination change can be solved with the advancement of the algorithm, but the difficulty of occlusion still needs long-term research.

References

1. Grobel K, Assan M. Isolated sign language recognition using hidden Markov models[C]. IEEE International Conference on Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation. IEEE, 2002:162-167 vol.1.
2. Reyes M, Dominguez G, Escalera S. Featureweighting in dynamic timewarping for gesture recognition in depth data[C]. IEEE International Conference on Computer Vision Workshops. IEEE, 2011:1182-1188.
3. Simo-Serra E, Ramisa A, Alenyà G, et al. Single image 3D human pose estimation from noisy observations[C] Computer Vision and Pattern Recognition. IEEE, 2012:2673-2680.
4. Sinha A, Choi C, Ramani K. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4150-4158.
5. Koren Y. The bellkor solution to the netflix grand prize[J]. Netflix Prize Documentation, 2009.
6. Zhu Yuanxin, Xu Guangyou, Huang Huangyou, etc. Dynamic isolated gesture recognition based on appearance[J]. Journal of Software,2000,11（1）：54-61.
7. Xiao Ling, Li Renfa, Zeng Fanzi, Qu Weilan. Dynamic gesture recognition method based on learning sparse representation[J]. Journal on Communications,2013,34（6）:128-135.
8. Liu Shuping, Liu Yu,Yu Jun, Wang Zengfu. Hierarchical static gesture recognition combining finger detection and HOG feature[J]. Journal of Image and Graphics,2015,20(6) : 0781-0788 .
9. Wang Kai, Yu Hongyang, Zhang Ping. Real-time gesture recognition based on AdaBoost algorithm and optical flow matching[J]. Microelectronics & Computer. 2012.5(29):138-141.
10. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceeding s of the IEEE conference on computer vision and pat tern recognition. 2015: 3431-3440.
11. Badrinarayanan V, Handa A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling[J]. arXiv preprint arXiv:1505.07293, 2015.
12. Oberweger M, Lepetit V. DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation[J]. 2017:585-594.
13. Zimmermann C, Brox T. Learning to Estimate 3D Hand Pose from Single RGB Images[J]. 2017:4913-4921.
14. Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.
15. Oberweger M, Wohlhart P, Lepetit V. Hands Deep in Deep Learning for Hand Pose Estimation[J]. Computer Science, 2016.
16. Li Qing, Tang Huan, Chi Jian Nan, Xing Yongyue, Li Huatong. Study on the method of gesture segmentation based on the improved maximum category variance method [J]. Acta Automatica Sinica, 2017, 43(4):528-537
17. Mao Yanming, Zhang Liliang. Gesture segmentation and recognition based on Kinect depth information[J]. Journal of System Simulation, 2015, 27(4):830-835
18. Zhang Yi, Yao Yuanyuan, Luo Yuan. HMM dynamic gesture recognition algorithm based on B parameter improvement[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition). 2015(10)：416-419.
19. Wang Xiying, Dai Guozhong, Zhang Xiwen, et al. Complex dynamic gesture recognition based on HMM-FNN model[J]. Journal of Software,2008,19（9）：2302-2312.
20. Zhang Yi, Wu Shihai, Luo Yuan. The method and application of signal track recognition based on

HMM[J]. Electro-Optic Technology Application. 2015（8）：650-656.

21. Xu Jiabin. Research on dynamic gesture recognition based on parallel HMM[D]. Beijing University of Posts and Telecommunications. 2016.
22. Yu Meijuan, Ma Xirong. Improvement of dynamic gesture recognition based on HMM method [J]. Computer Science.2011(1):251-252.
23. Wang Wanliang, Yang Jingwei and Jiang Yibo. Gesture recognition based on motion sensor[J]. Chinese Journal of Sensors and Actuators,2011,24(12)：1723-1727.
24. Mueller F, Bernard F, Sotnychenko O, et al. GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB[J]. 2017.
25. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversrial Nets[C].International Conference on Neural Information Processing Systems.2014,3:2672-2680.
26. Spurr A, Song J, Park S, et al. Cross-modal Deep Variational Hand Pose Estimation[J]. 2018.
27. Kendall A, Grimes M, Cipolla R. Convolutional networks for real-time 6-DOF camera relocalization. CoRR abs/1505.07427 (2015)[J]. 2015.
28. Xue junjie and Chen jianqiang. Research and implementation of weighted DTW gesture recognition method[J]. Information Technology.2015(11):125-129.
29. Yu Chao, Guan Shengxiao. Dynamic gesture tracking recognition based on TLD and DTW[J]. Computer Systems & Applications. 2015.10(24):148-154.
30. Li Kai, Wang Yongxiong, Sun Yipin. An improved DTW dynamic gesture recognition method[J]. Journal of Chinese Computer Systems. 2016.7(7):1600-1603.
31. He Chao, Hu Zhangfang and Wang Yan. A dynamic gesture recognition method based on improved DTW algorithm[J]. Digital Communication. 2013.6(40):21-25.
32. Zhou Zhiping and Miao Minmin. Combined DTW and improved STLCS dynamic gesture authentication[J]. Journal of Electronic Measurement and Instrumentation. 2015.7(29):1064-1073.