Sign language recognition based on depth image processing

Wanyu Zhang*

Department of Information and Mechatronic Engineering, Shanghai Normal University, Shanghai, China

*Corresponding author: zhangw33@udayton.edu

Abstract. Deaf and mute people mainly rely on sign language to communicate, but they still face many problems in their daily life. The interaction between humans and computers has developed extensively in recent years, the flexible and easy-to-understand characteristics of sign language make it suitable for application in the technology of computer vision. This paper mainly introduces sign language recognition from the current research status, application of deep image processing technology, collection of sign language datasets and the final results. After analysis, this paper finally established a recognition model with the wrist joint as the center point to complete the extraction of hand features. This paper input a large number of pictures to train the model, and the final recognition rate reached 91.3%. This paper can widely apply this recognition technology in computer-aided teaching, TV program bilingual broadcasting, virtual human research, medical research, game entertainment and many other aspects.

Keywords: Sign Language Recognition, Image Processing, Datasets

1. Introduction

The interaction between humans and computers has gradually become one of the most popular topics with the advent of the age of technology. Due to the limitations of language, there are still some communication barriers of deaf and mute people in many scenarios. Sign language, as the primary means of communication for deaf and mute people, is flexible and easy to understand. Therefore, sign language has broad application prospects in the technology of computer vision [1]. The second chapter in this paper introduces the current research status of sign language recognition, including research based on wearable devices, research based on computer vision, three implementation methods of sign language recognition technologies and the comparison of these three methods. The third chapter mainly introduces deep image processing techniques and how to apply deep image processing techniques to complete the gesture recognition. The fourth chapter mainly introduces the open source and self-collected sign language datasets. The fifth chapter analyzes and summarizes the results of sign language recognition. This paper extracted the gesture from the wrist joint and successfully extracted twenty-one 3D coordinate key points of the hand and obtained a high recognition rate. Sign language recognition has a far-reaching research significance and broad practical application prospects, because the use of gestures can improve the level of intelligence in VR. At the same time, this paper can further train and improve the computer's ability to understand human language and promote the development of the entire industry by studying the mechanism of human visual language understanding [2].

2. Research states of Sign Language Recognition

2.1. Research based on wearable devices

Wearable devices mainly rely on expensive equipment to extract and recognize gesture features of images. Its advantage is that it can still effectively extract gesture features in complex backgrounds, so this method has a high recognition rate [3].

For a long time, researchers have had in-depth research on wearable devices. The earliest data gloves were developed by RichSayre, who used the gloves in the field of system control. After AT&T in the United States improved the light-sensitive fiber inside, they launched the first generation of

CyberGlove gloves [4]. The glove was good at capturing the movement of gestures in space and detecting the curvature of the fingers. Since then, many researchers have started to research it. MW Kadous used the technology of the PowerGlove data glove to identify 95 Australian sign languages. Assam identified isolated sign language based on a hidden Markov model. They identified sign language videos recorded by experimenters wearing gloves of different colors, and the sign language recognition rate reached 91.3% [4].

Wearing gloves has the characteristics of high accuracy, but its price is very high, and the operability is relatively complicated, so wearing gloves cannot have a broad market for promotion. The model of the data glove is shown in Figure 1.



Figure 1. Data glove model

2.2. Research based on the technology of computer vision

Different from the research based on wearable devices, the research based on the technology of computer vision refers to the method of collecting gesture images based on cameras and then identifying them. There are two types of cameras, one is a Kinect camera and the other is a normal camera. The Kinect camera is good at capturing color images and processing depth data at the same time [5]. However, the Kinect camera is more limited by environmental requirements, because it is greatly affected by light, so compared with ordinary cameras with good stability, its market is more severely limited. The performance of ordinary cameras is stable, and it is easier to introduce to the market.

Sign language recognition based on computer vision generally includes several steps, it contains segmentation of gesture images, extraction of gesture features and image recognition. Ordinary cameras can only extract two-dimensional images of gestures, so the input image needs to be segmented, and the segmented graphics need to be preprocessed to obtain the extraction features of the target image.

Starner used HMM-related recognition technology and collected hand centroid features through ordinary cameras. The recognition rate reached 87.8% [6]. Based on the ANN classifier, Kishore used Fourier correlation technology to collect gesture images through four cameras, effectively solving the problem of hand occlusion. The final recognition rate reached 90.2% [7]. Rashid used geometric features and Zernike moments as feature vectors, and used SVM sign language recognition technology to recognize finger language in American Sign Language, and the recognition rate reached 98.5% [8].

2.3. Research on the methods of Sign Language Recognition

For the technology of sign language recognition, the implementation methods are mainly divided into their aspects:

(1) Recognition based on geometric features. This method uses the edge features and regional structural features of gestures as the recognition features. The benefit is that it is adaptable and stable in gesture recognition, but the learning ability of this method is relatively low and the recognition rate will not improve significantly when the sample size continues to increase.

(2) Based on deep learning and neural network recognition. Due to the complex nonlinear nature of the method, it is not susceptible to interference. This technology is always used to complete the recognition of static gestures, but its disadvantage is that the single-layer neural network has only a low feature learning ability, which will lead to the situations of overfitting and local optimum.

(3) Recognition based on Hidden Markov Model (HMM). This model has a strong descriptive ability, and it can accurately control the changes of gesture signals. However, this model has a low recognition rate because it requires a lot of calculation of parameters.

3. Gesture recognition based on depth image processing technology

3.1. Palm-centered apparent feature extraction

By extracting the apparent features of the gesture, the gesture is classified according to the number of fingers representing the gesture and the angle between the fingers, so as to recognize the gesture of rotation and scaling. Compared with other gesture feature extraction methods, the apparent features of gestures have the advantages of being more intuitive, requiring no training samples, a strong adaptability, a fast-computing speed, and a good real-time performance.

(1) Appearance feature extraction steps of gestures

① Obtain the center point of the gesture area through the erosion operation in mathematical morphology. Since the palm is the main component of the gesture surface, it occupies the largest area in the gesture area. The boundary points of the gesture area can be eliminated by using the techniques of corrosive operations, and then the gesture area can be gradually reduced. Finally this paper can obtain the center point C_0 of the gesture area. Figure 2 shows the actual effect of the extraction of the center point.



Figure 2. The extraction of the center point

② There is some distance between the center point and the edge of the gesture area. Calculate the maximum distance and mark it as 1. Then divide the distance into 10 equal parts, it means that d=l/10. Take the center point of the gesture area as the center of the circle to make a circular trajectory, the radius of the circle starts from d to l, and the value of d is incremented each time, and this paper can obtain 10 circular trajectory lines as shown in Figure 3.

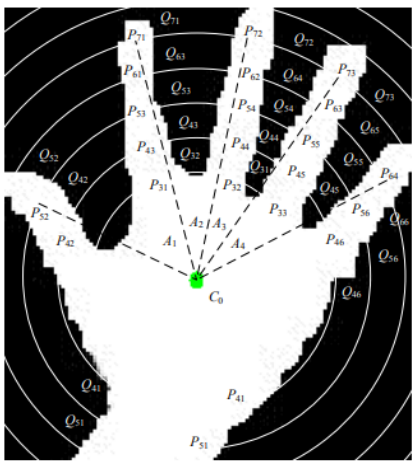


Figure 3. Circular trajectory lines of hand

③ Record the position coordinates of the pixel value change points p_{ij} and Q_{ij} on each circular track in a clockwise direction value. i refers to the number of the trajectory circle, j refers to the number of the P or Q point on the same trajectory circle, and delete the independent p_{ij} point and Q_{ij} point at the same time.

④ Calculate the distance D_{ij} between each pair of p_{ij} and Q_{ij} through the position coordinates of p_{ij} and Q_{ij}. When the trajectory circle intersects with the fingertip part of the area, a smaller D_{ij} can be obtained. When D_{ij} is less than the threshold δ, delete the corresponding p_{ij} and Q_{ij} points. It can be seen that the p_{ij} and Q_{ij} points are removed from the interphalangeal part of some fingers, and the threshold is set to δ=d/4 according to experience.

⑤ The maximum value of j obtained on each trajectory circle is the sum of the number of branches connected to the palm. The mathematical expression is $N = \max(j)$. Since the branch contains fingers and wrist branches, the number of fingers can be obtained by subtracting 1 from the sum of the number of branches. The mathematical expression is $N_f = N - 1$.

⑥ Calculate the average width of the branch from the average value of each branch. The mathematical expression is $W_j = D_{ij}$. It can be seen from the hand shape that the width of the wrist is larger than the fingers, and in the branches, the wrist corresponds to the branch with the largest width. The middle point of p_{ij} and Q_{ij} is connected with the center position point C_0 on the maximum trajectory circles intercepted by other branches except the wrist branch, and the angle A_{j-1} between the fingers can be obtained.

(2) Rotation and scaling invariance

Human-computer interaction requires natural gestures, not limited to certain specific gesture poses or specific gesture area sizes. The distance and position of the gesture can cause problems with the size and rotation of the gesture in the image. The feature extraction method in this paper uses the equally divided trajectory circle to eliminate the influence of the size of the gesture area on the feature extraction. And at the same time, this method also calculates the features of the changing points on the trajectory. The extracted finger number N_f and inter-finger angle A_{j-1} features have rotation and scaling invariance, and they will not be affected by gesture distance and rotation.

3.2. Feature extraction from wrist joint

Based on the existing computer vision technology, this paper choose a finger tracking solution called MediaPipe Hands with the principle of high precision and multi-dimensionality as possible. As one of the important methods used in artificial intelligence systems, machine learning (ML) endows the hand tracking process with the ability to learn, reason, perceive and solve problems. Mediapipe uses ML to extract and analyze a single frame to infer the twenty one 3D landmarks of the hand.

The ML pipeline consists of multiple models, which cooperate with each other to complete the scanning, calculation and display of the hand: one is the palm detection model, its role is to scan the entire palm, and according to a certain direction, display the scanned bounding box of the hand; the second is the hand landmark model. The image becomes a bounded image after the palm detection model, then the hand landmark model operates and calculates on these images, thereby inferring the values of 21 key points of the hand, and transforming to display a high-quality and high-fidelity 3D map of the hand.

The hand landmark model provided by this method can realize fast scanning, fixed-point cropping, accurate calculation of hand images, and does not require special image processing, such as up and down rotation, left and right translation, zooming in and out, and so on. Most of its functions can be used to predict the accuracy of coordinates. In addition, in the pipeline, the landmark model has a

higher priority than the palm detection model, and only when the landmark model cannot identify the hand, the palm detection will be used instead to relocate the hand.

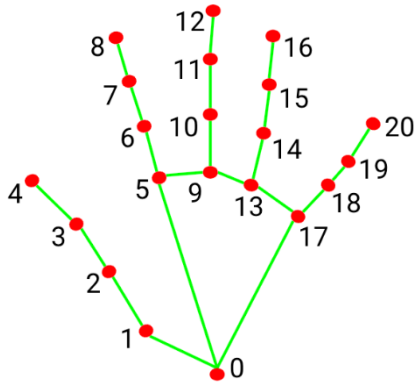
(1) Palm detection model

This paper first need to design a one-shot hand detector model. The main purpose is to detect the initial position of the hand. The model has an optimization feature that reduces the effects of real-time hand motion. Since the size of each detected hand is different, it is required that the image frame of this model must have a larger span, at least 20 times, in order to process hand images of various sizes. In addition, the model is also required to detect hands in two special cases, one is an occluded hand, and the other is a self-occluded hand. Hand recognition technology is more difficult than many current facial recognition technologies on the market, because the eye, mouth and face contour area of the face are all high-contrast patterns in the frame images, while the hands lack these features, so it is relatively difficult for the computer to detect them reliably. Therefore, providing more information related to the hand, such as features of arm, body, gender, etc., can scan and detect the position of the hand more accurately.

This paper need to train a hand detector instead of a palm detector first. Due to the rigid features of the palm and the fist, it is characterized by obvious boundaries, but the fingers have multiple flexible joints, which control the bending and rotation of the fingers, so it is not easy to detect. The size of the palm is small, and this paper can use the non-maximum suppression algorithm to complete the self-occlusion of hands such as making a fist. Furthermore, ML refers to square bounding boxes as anchors, and palms can be modeled by using anchors. The benefit is that other aspect ratios of the hand can be ignored and the number of anchors will be reduced by a factor of 3–5.

(2) Hand landmark model

After detecting the entire presented palm image, this paper use twenty one 3D hand joint coordinates to accurately locate the important key points of the hand landmark model. The benefit of this model is that it is sturdy even if the hands are partially visible. Figure 4 shows the twenty one 3D coordinates of the hand.



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figure 4. Twenty one key points of the hand

4. Collection of sign language datasets

There are two main types of datasets, open source datasets and self-collected datasets.

4.1. Open Source Datasets

Open source datasets include the following:

① PHOENIX-Weather continuous sign language datasets [9]

They contain 45,760 video samples provided by 9 people, including 5,356 sentences related to weather forecasts.

② SIGNUM datasets

There are 33,210 video samples in the datasets. They were provided by 25 people. 780 sentences and 450 German Sign Language words can be found, and each sentence contains 2 to 11 Sign Language words.

③ Boston University's ASLLVD datasets

They contain 9800 video samples provided by 6 signers based on over 3300 American Sign Language vocabulary.

④ ASL-LEX American Sign Language datasets

They contain various features of nearly 1000 sign languages.

In addition to sign language datasets, there are sign language alphabet datasets, gesture shape datasets, data glove-based sign language datasets, and datasets made with other specific wearable devices [10].

4.2. Self-collected datasets

In daily research, in addition to researching open source datasets, self-collected images are also very important as training samples.

4.2.1 Requirements

① Requirements for image acquisition equipment: ordinary camera.

② Richness of images: the same gesture, different backgrounds, different lighting, different angles, and different people are considered comprehensively; sign language characteristics of different gestures.

4.2.2 Collection of valid samples

It is not necessary to collect as many samples as possible. The complexity of the samples will cause overfitting during image classification and recognition; therefore, in the image pre-classification stage, screening samples is the most critical step.

The stage of picking feature samples:

(1) First, the selected samples need to contain specific gesture features, and these gestures can represent specific sign language meanings. This is the most basic requirement to complete sample collection.

(2) In the process of selection, it is necessary to avoid sample duplication, which is to ensure the diversity of samples.

(3) The samples need to be collected from multiple angles, which is to eliminate the interference of external equipment.

(4) The set of selected samples can represent a motion state of a specific hand shape. The purpose is to ensure the effective operation of the design model.

(5) For a one-handed characteristic hand shape, the selected samples need to contain different cases of left and right hands.

(6) The hand shape with the arms hanging down naturally is also one of the characteristic hand shapes. It refers to the starting state.

4.2.3 Method of application

By extracting the apparent features of gestures, this paper can establish a decision tree model [10] to recognize and classify gestures. Decision tree model is a mathematical classification method. When this paper input the samples which need to be trained in the decision tree model, the model can generalize and learn from these samples and then form a specific set of rules. These rules are called decision rules. The decision tree model then uses these rules to classify the new data. There are two types of nodes in the decision tree, one is called the root node and the other is called the leaf node. A decision tree arranges the instances from the root nodes to the leaf nodes to complete the classification of these instances. The leaf node where the instance is located finally is its corresponding category. When establishing a decision tree model, this paper need to use a certain logical judgment ability to design it, which is the decisive factor in judging whether the decision tree can be successful. The

main extraction features of gestures are the number of fingers N_f and the angle A_{j-1} between the fingers, which are used as classification nodes of the decision tree. The number of fingers between different gestures is clearly distinguished, as shown in Figure 5.

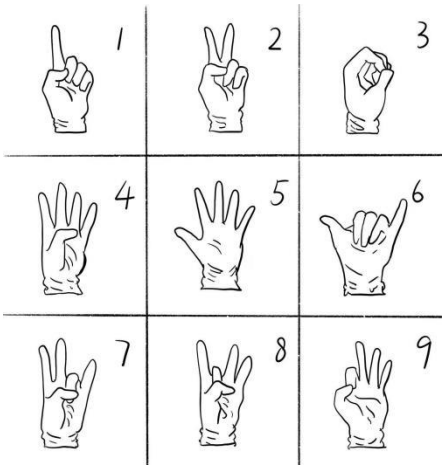


Figure 5. Different gestures

The established decision tree first takes the number of fingers N_f as the root node of the decision tree. For gestures with the same number of fingers, it then establishes sub-nodes to distinguish the difference of angles A_{j-1} between the fingers. For the angles A_{j-1} between the fingers on the sub-nodes, the classification conditions can be obtained through the test training of small samples, and the established decision tree is shown in Figure 6.

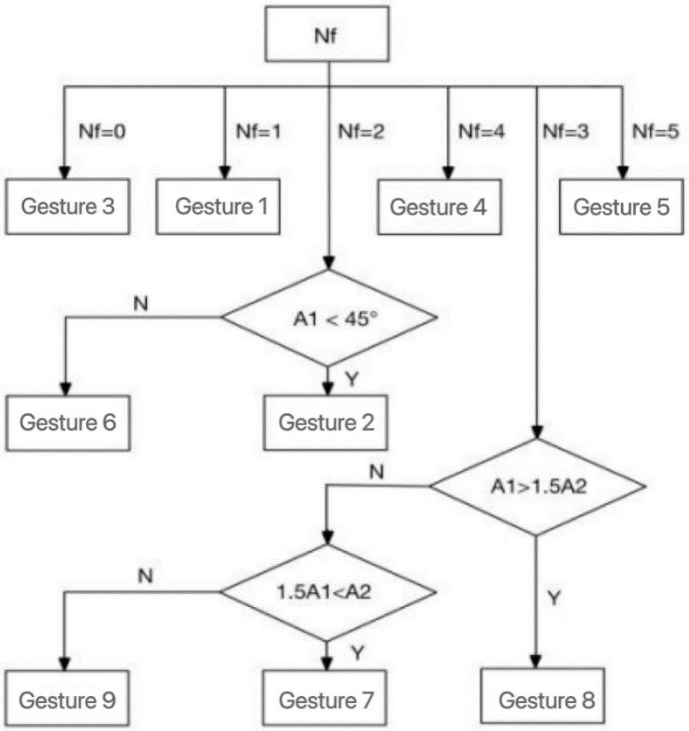


Figure 6. Gesture classification decision tree

The number of fingers of gesture 1, gesture 3, gesture 4, and gesture 5 is unique and can be directly classified by the root nodes; N_f = 2 for gesture 2 and gesture 6, they can be distinguished by judging the size of the angle A1 between the fingers through the next-level sub-node; N_f = 3 for gesture 7, gesture 8, and gesture 9, and they can be distinguished by judging the size of the angle A1 and A2 between the fingers through the two layers sub-nodes.

5. Results

5.1. The extraction of twenty one 3D key points of hand

After comparison and analysis, this paper finally chose to use the wrist joint as the starting point to extract the characteristics of the hand. This paper finally successfully realized the extraction of twenty one 3D key points of hand by writing code through Pycharm and using the Mediapipe program. The extraction result is shown in Figure 7.

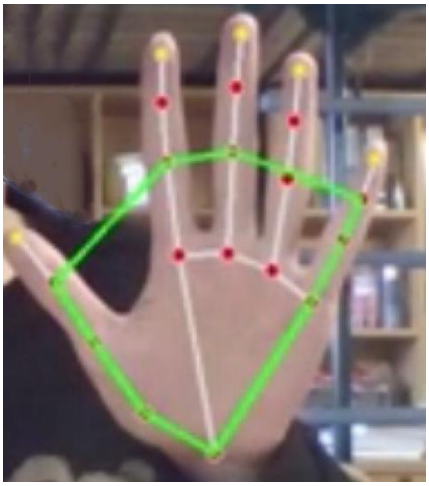


Figure 7. Twenty-one 3D key points of hand

5.2. Analysis of experimental results

After successfully extracting 21 key points of the hand, this paper input a large number of sign language pictures, and use image processing technology to extract the key features of the hand, so as to train the model and test the recognition accuracy of the model. This paper trained 54 gestures in total, including 1-9 numbers, good, bad, I love you and other common words. In the first recognition, the model accurately identified the meaning of 49 words. In the second recognition, the model accurately identified the meaning of 48 words. In the third recognition, the model accurately identified the meaning of 51 words. Table 1 shows the results of three tests.

Table 1. Test of wrist joint model

Number of test	Total	Successful identification	Recognition rate
1	54	49	90.7%
2	54	48	88.9%
3	54	51	94.4%

Through the table, the final recognition rate of the model is 91.3% by averaging the recognition rate obtained from the three tests. From this, it can be seen that a high recognition rate can be obtained by recognizing gestures based on depth image processing technology, and this method can be widely promoted.

6. Conclusion

This paper mainly introduces sign language recognition based on deep image processing technology through four aspects. The first part researches on the current status and three implementation methods of sign language recognition. Then the second part introduces how to apply deep image processing technology in gesture recognition. There are two methods of gesture feature extraction, one is starting from the palm center and the other is starting from the wrist joint. The third part focuses on the collection of sign language datasets. The datasets mainly include open source datasets and self-collected datasets. The results of sign language recognition are presented in the last

section. The development of this technology makes the interaction between humans and machines more intelligent, and makes more companies pay attention to the development of artificial intelligence, thereby promoting the further expansion and barrier-free development of the entire artificial intelligence industry, and solving more problems.

References

[1] Lindsay Hinzman and Spencer D. Kelly. Effects of emotional body language on rapid out-group judgments[J]. Journal of Experimental Social Psychology, 2013, 49(1) : 152-155.

[2] Starner T , Massachusetts Institute of Technology. Media Laboratory. Vision and Modeling Group. Visual Recognition of American Sign Language Using Hidden Markov Models[J]. proc fg, 1995.

[3] Qiqi Shi. Research on Sign Language Recognition Based on Improved Convolutional Neural Network [D]. Xi'an University of Architecture and Technology, 2019. DOI: 10.27393/d.cnki.gxazu.2019.000290.

[4] Kostas N. Tarchanidis and John N. Lygouras. Data Glove With a Force Sensor[J]. IEEE Transactions on Instrumentation and Measurement, 2003, 52(3) : 984-989.

[5] Xiao-Li Guo and Ting-Ting Yang. Gesture recognition based on HMM-FNN model using a Kinect[J]. Journal on Multimodal User Interfaces, 2017, 11(1) : 1-7.

[6] Starner A T . Visual Recognition of American Sign Language Using Hidden Markov Models[C]// Proc IEEE Intl Workshop on Automatic Face & Gesture Recognition. 1995.

[7] Kishore P . 4-Camera model for sign language recognition using elliptical fourier descriptors and ANN[C]// International Conference on Signal Processing & Communication Engineering Systems. IEEE, 2015.

[8] Ahmed O R , Al-Hamadi A , Michaelis B . Utilizing Invariant Descriptors for Finger Spelling American Sign Language Using SVM[C]// Advances in Visual Computing - 6th International Symposium, ISVC 2010, Las Vegas, NV, USA, November 29-December 1, 2010. Proceedings, Part I. Springer-Verlag, 2010.

[9] Paul, Soumi et al. Development of benchmark datasets of multioriented hand gestures for speech and hearing disabled[J]. Multimedia Tools and Applications, 2022, : 1-37.

[10] Gaolin Fang and Wen Gao and Debin Zhao. Large Vocabulary Sign Language Recognition Based on Fuzzy Decision Trees[J]. IEEE Transactions on Systems Man and Cybernetics Systems, 2004, 34(3) : 305-314.