Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/141/2025.21696

Vision-based Hand Gesture Recognition Technology

Xinyue Lou[1,a,*]

[1]*Zhejiang A&F University, College of Chemistry and Materials Engineering, Hangzhou, 311300, China*
a. xy_Lou0121@163.com
**corresponding author*

Abstract: Human-computer interaction has a wide range of application prospects in many fields such as medicine, entertainment, industry and education. Gesture recognition is one of the most important technologies for gesture interaction between humans and robots, and visual gesture recognition increases the user's comfort and freedom compared with data glove recognition. This paper summarizes the general process of visual gesture recognition based on the literature, including three steps: pre-processing, feature extraction, and gesture classification. It also defines static and dynamic gestures and makes a comparison between their differences and recognition emphases. Based on static and dynamic gesture recognition, this paper summarizes the commonly - used visual gesture recognition methods. For static gesture recognition, it includes methods such as the template - matching method and the AdaBoost - based method. As for dynamic gesture recognition, it encompasses methods like the hidden Markov model method and the dynamic time regularization method. Finally, some applications of visual gesture recognition are introduced, for example, a non-contact system for operating rooms and smart home control.

Keywords: gesture recognition, posture recognition, computer vision, human-computer interaction

1. Introduction

People interact with various electronic devices more and more in daily lives as intelligent and electronic technology grows in popularity. Gesture interaction changes human-computer interaction from the traditional machine-centric mode to the human-centered one[1], which conforms to the trend of naturalizing human-computer interaction. As an important part of robot-gesture interaction, the accuracy and rapidity of gesture recognition directly impact the accuracy, fluency, and naturalness of human-robot interaction[2]. Since the 1970s, researchers have made a great many research achievements in the field of gesture recognition and interaction technology. However, researchers and users are not satisfied with the recognition technology based on data gloves, they hope that the interaction can be richer, convenient and efficient, so the gesture recognition technology based on vision has begun to develop. Compared with contact gesture recognition, vision-based gesture recognition does not require wearable devices, and the interaction is more convenient and flexible. It has a wide range of applications and becomes a current hot topic[3]. Based on the literature, this paper summarizes the general process of visual gesture recognition, explains the definitions of static and dynamic gestures, compares their differences and recognition emphases, and sums up the commonly-

DOI: 10.54254/2755-2721/141/2025.21696

used visual gesture recognition methods. Finally, it introduces some applications of visual gesture recognition.

2. Steps for Visual Gesture Recognition

Vision-based gesture recognition mainly consists of three steps: pre-processing, feature extraction, and gesture classification. The specific steps are shown in Figure 1. Pre-processing includes image acquisition, gesture detection, and segmentation[4]. First, the camera captures the image or video containing the human hand. Second, the hand and its specific position in the image or video are detected and separated from the background[2], which involves the extraction and fusion of interactive information such as color, shape, and movement. The effect of gesture segmentation is directly related to the accuracy of gesture recognition[5]. Feature extraction includes the extraction of global features and local features of the hand. Global features include color, texture, shape, etc., and local features include corner classes and region classes. In the gesture classification step, model construction and evaluation are carried out. This is to determine the accuracy of the built model and adjust it until it meets expectations. Then, gestures are classified by comparing the extracted features with the model. Finally, the recognition results are sent to the computer for interactive feedback[4]. After receiving feedback from the computer, the user give feedback to the camera to better meet the user's needs and form a closed loop.

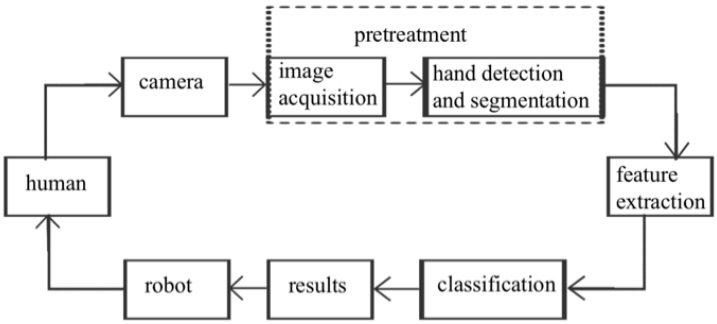

Figure 1: Visual Gesture Recognition Flow Diagram

3. Methods of Visual Gesture Recognition

Based on temporal relationships, gestures can be divided into static gestures (hand postures) and dynamic gestures (hand gestures), as shown in Figure 2. The gesture image is regarded as the research object for static gestures. During the gesturing period, the position of the hand remains unchanged and corresponds to the spatial point of the model parameters. The research object of dynamic gesture is the gesture video, and the position of the hand changes continuously with time, corresponding to a trajectory in the parameter space of the model. The model for static gesture recognition is mostly based on gesture image attributes, such as structure, boundary, image feature vector, area histogram feature, etc. While for dynamic gesture recognition, the model is mainly based on the change or motion trajectory of the image. Dynamic gestures are much higher than static gestures in complexity and the amount of information they contain, because dynamic gestures are equivalent to a series of static gestures. The recognition of dynamic gestures not only requires the static gestures in each frame to be separated from the background and feature extracted, but also classified according to the order of gestures, combined with temporal and spatial features.[2] [4] [6]

DOI: 10.54254/2755-2721/141/2025.21696

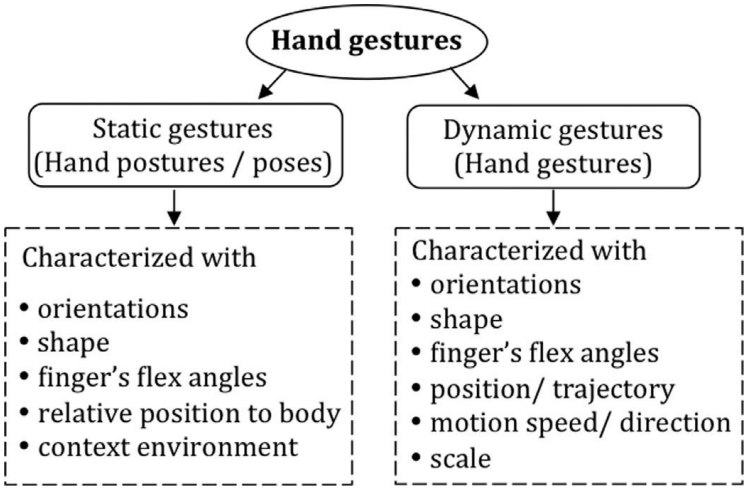

Figure 2: Classification of hand gestures.[7]

The following is an introduction to gesture recognition according to the classification of static and dynamic gestures, classifying one category does not mean that it cannot be used for the recognition of another type of gesture, but it is more or more suitable for that type of gesture.

3.1. Static Gesture Recognition

3.1.1. Template Matching-based Methods

One of the oldest and most basic pattern recognition techniques is the template matching method, which is mostly employed for static gesture recognition. The method is to match the input image with the template (point, curve or shape) and classify it according to the similarity of the match. The advantages of the template matching technology include simplicity, speed and resistance to lighting, background, posture, etc., and has a wide range of applications. However, the classification accuracy is low, and the types of gestures that can be recognized are limited. Therefore it is suitable for small samples and factors such as shapes do not change much.[2]

3.1.2. AdaBoost-based Methods

The boosting algorithm is a statistical learning method that promotes the weak learning algorithm to the strong learning algorithm. A series of basic classifiers (weak classifiers) are constructed by repeatedly adjusting the weight distribution of the training data. These basic classifiers are then linearly combined to create a strong classification. AdaBoost offers some benefits: AdaBoost provides a framework within which various methods can be used to construct subclassifiers. Without filtering characteristics, it can employ a basic weak classifier, and overfitting rarely occurs. The classification accuracy of every weak classifier determines the accuracy of the strong classifier produced by AdaBoost. AdaBoost can adaptively adjust the assumed error rate in response to the weak classifier's output, which is very efficient and can significantly improve learning accuracy. However, during the training process, AdaBoost makes the weights of the samples that are difficult to classify increase exponentially, causing the training to be overly biased towards these challenging samples. This has an impact on error computation and classifier selection, lowering the accuracy of classifiers. Furthermore, AdaBoost is prone to noise interference, and its execution effect depends on the selection of weak classifiers, which requires a lengthy training period.[2]

DOI: 10.54254/2755-2721/141/2025.21696

3.1.3. Artificial Neural Network(ANN)-based Methods

The artificial neural network is a widely parallel interconnected network made up of simple and adaptable units, which was born in the early 40s of the 20th century. Characterized by high fault tolerance and parallelism, robustness, adaptability, mobile learning and anti-interference abilities, it can simulate how the biological nervous system interacts with the outside world. Neural networks come in a variety of forms, and the quality of the hand detection model and the quantity of training samples often determine the gesture recognition rate.[2]

3.1.4. Geometric Features-based Methods

The method based on geometric features is a gesture recognition method that uses geometric features to judge the number of unfolded fingers. Since various gestures can be distinguished by the number of unfolded fingers, some researchers identify gestures by counting the number of unfolded fingers, which can be calculated by using the intersection of circles or lines and gesture contours.[2]

3.2. Dynamic Gesture Recognition

3.2.1. Hidden Markov Model (HMM)

The most frequently used hand gesture recognition (HGR) technique is the hidden Markov model[7]. The hidden Markov model is a probabilistic model about variable sequences in time, in which there are two random processes: the observable state and the hidden state. A random sequence of unobservable states is generated at random by a hidden Markov chain, which then creates an observation from each state to produce an observed random sequence. People often need to solve valuation problems, decoding problems, and learning problems when using HMMs[1]. The hidden Markov model is better suited for continuous gesture recognition, especially for complex gestures involving context. The computational effort of hidden Markov model training and recognition is very large, especially in the analysis of continuous signals. The state transition requires a large number of probability densities calculations, so the number of parameters rises, slowing down sample training and target recognition. The discrete hidden Markov model is used for analysis in the generic gesture recognition system to address this issue.[2]

3.2.2. Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is a nonlinear time normalized pattern matching technology that combines time regularization with distance gesture measure calculation. DTW algorithm involves two sequences: a template sequence and a sequence to be measured. Its main idea is to use dynamic programming to find a scientific time calibration matching path, which is the mapping relationship between points on the sequence. For two sequences with differences in the timeline, the DTW algorithm can eliminate the difference in the timeline and minimize the distortion between them, by expanding, compressing or deforming the sequence globally or locally, and the overlap with the timeline of the other sequence reaches the maximum[1], to eliminate the nonlinear fluctuation in time[2]. The DTW algorithm matches the gesture to be tested with each template in the template gesture library, and the template with the least DTW distance is returned as the recognition result.[1] The dynamic temporal regularization method is concise and has low training overhead. It can eliminate the temporal differences among various spatiotemporal modes, process temporal context, and is scalable. Thus, it is more suitable for dynamic gesture recognition problems.[6] However, the size of the sample bank limits this method, and the recognition impact and stability are subpar, especially when dealing with large numbers, complex gestures and combined gestures.[2]

DOI: 10.54254/2755-2721/141/2025.21696

3.2.3. Support Vector Machine (SVM)

Support vector machine (SVM) is a binary classification model, and its basic model is a linear classifier that defines the maximum interval in the feature space[2], which finds an optimal "line" in the feature space so that the two types of objects are distributed on both sides of the "line" as much as possible[6]. Using kernel approaches, SVMs can also be extended into nonlinear classifiers. Because the normalized central moment is sensitive to direction changes, and has translation and proportional invariance[2], SVM generally divides the samples into multiple classes by constructing multi-level features for binary classification iteration or mapping the samples to a higher-dimensional space and looks for hyperplanes for multi-classification[6].

4. Applications of Visual Gesture Recognition

Visual gesture recognition technology is developing towards higher accuracy, faster response speed, and a wider range of application scenarios. In healthcare, visual gesture recognition can be used for contactless interactions in a sterile environment, thereby reducing the risk of infection for patients. Gallo et al. introduced an open-source system for a highly interactive, controller-free exploration of medical images. Users can interact with the interface of the system at a distance using hand and arm gestures.[8] In addition, visual gesture recognition technology can provide rehabilitation treatment and convenient living equipment for people with disabilities. For example, it can translate sign language into speech or text, allowing deaf people to communicate more smoothly with people with normal hearing. Computers can understand the information transmitted by dynamic human gestures through gesture detection and recognition technology, and then further translate the information into different types of text or speech through translation technology. In this way, these technology help people with hearing or speech impairment communicate, or help people with learning disabilities interact with computers[9]. It can also be combined with assistive devices such as electric wheelchairs, and users can control the direction and speed of movement of the wheelchair through gestures, improving the autonomy of travel.

At home, people can control smart home appliances such as lights and TVs more easily by applying visual gesture recognition, to achieve a more convenient smart home experience. Besides, gesture detection and recognition technology can be used to help humans perceive dangerous events and abnormal behaviors, and provide alarms to notify guardians in time[9].

In terms of game entertainment, it is not only possible to bring people an immersive gaming experience by incorporating virtual reality technology, but also to manage the entertainment system. Mendels et al. developed a user identification system. By completing many motion signatures, users trained the system and interactively developed their own motion signatures through demonstration. The proposed system can be utilized for parental control, content adaptation, and interface customization when incorporated into gesture-based home entertainment systems.[10]

In the field of education, visual gesture recognition supports interactive teaching and provides more vivid and rich teaching content. In the automotive industry, drivers can use gestures to control navigation and audio systems without the distraction of looking for physical buttons or touching the screen. Ohn-Bar and Trivedi developed an In-Vehicle Gestural Interface based on instantaneous vision. It offers drivers a lower visual load, fewer driving errors, and a higher degree of user acceptability than touch interfaces. It can handle the distraction from today’s increasingly complicated function in-vehicle interfaces[11].

5. Conclusion

This paper first presents the steps of visual gesture recognition, which mainly include pre-processing, feature extraction and gesture classification. Pre-processing encompasses image acquisition, hand

DOI: 10.54254/2755-2721/141/2025.21696

detection and segmentation; feature extraction involves extracting global features and local features; and gesture classification consists of model construction, model evaluation and feature comparison. In addition, it has shown that the research object of static gesture is gesture image, while the research object of dynamic gesture is gesture video. Dynamic gesture should be classified according to the order of gestures, combined with time and space characteristics based on static gestures. Furthermore, the concept, principle, advantages, and disadvantages of visual gesture recognition methods for both static and dynamic gestures are introduced. Finally, it presented some applications of visual gesture recognition, which involve various fields like medical health, games and entertainment, education, etc. For example, there are non-contact systems in operating rooms, and in-car navigation and audio control systems. Since no empirical research is used and the number of literature used is limited, this paper may not be comprehensive. Researchers can conduct empirical research and analyze more literature in future studies.

References

[1] *Wu. , Xie & Zhou. (2016). Research on human-computer interaction technology based on gesture recognition. Computer age (2), 4.*

[2] *Qi. , Xu & Ding . (2017). Research progress of robot visual gesture interaction technology. Robot (04), 565-584. DOI: 10.13973/J.CNKI.robot.2017.0566.200106006066*

[3] *Wei & Wang. (2024). Overview of gesture recognition and interaction. Computer and modernization (08), 67-76.*

[4] *Zhang et al. (2024). Research on gesture recognition technology and methods. Journal of Northwest University for Nationalities (Natural Science Edition) (02), 21-36. DOI: 10.14084/j.cnki.cn62-1188/N.*

[5] *Tian, Yang , Liang & Bao. (2020). Overview of Visual Dynamic Gesture Recognition. Journal of Zhejiang Sci-Tech University (Natural Science Edition) (04), 557-569.*

[6] *Fu , Li & Luo. Overview of dynamic gesture recognition technology based on vision. Computer Measurement and Control 1-15.*

[7] *Pisharady, P. K. , & Saerbeck, M. . (2015). Recent methods and databases in vision-based hand gesture recognition: a review. Computer Vision and Image Understanding, 141(C), 152-165.*

[8] *Gallo, L., Placitelli, A. P. , & Ciampi, M. . (2011). Controller-free exploration of medical image data: Experiencing the Kinect. CBMS '11: Proceedings of the 24th IEEE International Symposium on Computer-Based Medical Systems. IEEE.*

[9] *Zhang (2023). Research on dynamic human gesture detection and recognition technology based on computer vision (Ph.D. dissertation, Beijing University of Technology). Dr. https://link.cnki.net/doi/10.26935/d.cnki.gbjgu.2023.000367doi:10.26935/d.cnki.gbjgu.2023.000367.*

[10] *Mendels, O. , Stern, H. , &Berman, S. . (2017). User identification for home entertainment based on free-air hand motion signatures. IEEE Transactions on Systems Man & Cybernetics Systems, 44(11), 1461-1473.*

[11] *Ohn-Bar, E. , & Trivedi, M. M.(2014). Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. IEEE Transactions on Intelligent Transportation Systems, 15(6), 2368-2377.*