Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/43/20230810

A comprehensive analysis of gesture recognition systems: Advancements, challenges, and future direct

Shijia Li[1,5,9,†], Luoyun Zhou[2,6,†], Mingqi Fan[3,7,†], Yucheng Xiong[4,8,†]

[1]Software Engineering, Chengdu University of Technology, Chengdu, 610059, China
[2]Aritificial Intelligence, Jilin International Studies University, Changchun, 130117, China
[3]Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China
[4]Guangdong Experimental High School International Department (AP), Guangzhou, 528306, China

[5]sk1182199663@gamil.com
[6]zlandh2023@outlook.com
[7]fanmingqi0507@gmail.com
[8]yuchengxiong@hotmail.com
[9]Corresponding author email: sk1182199663@gamil.com
[†]Shijia Li, Luoyun Zhou, Mingqi Fan, and Yucheng Xiong contributed equally to this work and should be considered co-first authors.

Abstract. Gesture recognition emerges as a potent avenue for human-computer interaction, harnessing mathematical algorithms to interpret gestures. It promises to surpass text-based or graphical interfaces, enabling touchless device control through simple gestures. Our review of 7 papers encompassing various fields and methods underscores its diverse applications. Challenges persist, such as distinguishing genuine user intent from accidental actions amid environmental interference. Creating a universal EMG pattern recognition model demands intricate individual pre-training. Sensor-based gesture recognition grapples with real-world dynamics, necessitating adaptable models that discern user intent from non-intent actions. Addressing these gaps holds the key. Adaptable models and personalized approaches can enhance robustness and accuracy across applications, surmounting challenges in the gesture interaction technology realm.

Keywords: Gesture recognition, Wearable sensors, Machine learning, Human-computer interaction, Real-world dynamics

1. Introduction

Gesture recognition is a crucial element of human-computer interaction, allowing users to engage with devices and systems through natural gestures. A significant focus within this field involves the utilization of noninvasive wearable sensors for recognizing hand gestures.

This article presents a comprehensive review of contemporary methods for hand gesture recognition using wearable sensors. It explores machine learning techniques like support vector machines (SVMs)

DOI: 10.54254/2755-2721/43/20230810

and artificial neural networks (ANNs), as well as deep learning methodologies such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [1] [2]. The applications of gesture recognition using wearable sensors span various domains, including healthcare, gaming, and human-robot interaction.

The utilization of CNNs primarily revolves around analyzing visual data, while LSTM networks excel at capturing long-term dependencies and patterns in sequential data, proving useful in domains such as natural language processing and speech recognition. Additionally, the fusion of the residual model and the convolutional short-term memory model in ResNet enhances the extraction of essential features.

The paper also addresses two categories of challenges. One pertains to interference signals arising from redundant channels when using general-purpose equipment for fine movements. Solutions include a multi-stream CNN framework, simultaneous utilization of surface electromyography (sEMG), and the integration of linear discriminant analysis (LDA) and extreme learning machine (ELM) techniques [3]. Another challenge involves accurate gesture recognition within specific observation latencies, leading to the presentation of corresponding solutions.

Strengths and limitations of various approaches are discussed, alongside an acknowledgment of challenges in VR gesture recognition, encompassing real-time recognition, occlusion, lighting, and user-friendly interfaces.

In summary, the article offers a comprehensive overview of gesture recognition technology, emphasizing the potential of wearable sensors and VR environments. It details methodologies, applications, and insights into future directions, structured around a framework involving gesture recognition's basic process, problem-solution analysis, dataset requirements, research gap identification, and application scenario analysis.

2. Literature review

2.1. Computer Vision-Based Gesture Recognition

The computer vision-based limb recognition system encompasses three primary modules: hand detection, gesture recognition, and recognition-driven human-computer interaction (HCI) [4].

The hand detection module serves as the initial component of the system and is responsible for localizing hands within images captured by a monocular camera. At this stage, a skin color model is employed to effectively segment the hands from the background, accurately delineating the hand regions.

Subsequently, the gesture recognition component comes into play, with the task of identifying gestures from the hand images acquired by the camera. To achieve this, a Convolutional Neural Network (CNN) is employed as the primary tool to directly recognize gestures from the grayscale input of hand images. Xu(2017)stated the CNN model utilized has been trained on a dataset comprising 3,200 gesture images, encompassing 10 distinct gestures, each demonstrated by 10 different individuals, resulting in a total of 100 gesture samples per gesture [5].

Ultimately, the human-computer interaction (HCI) module is responsible for translating the recognized gestures into specific mouse or keyboard events. To realize this, a predefined mapping table is utilized to map each recognized gesture to corresponding mouse or keyboard events. The design of this mapping table aims to achieve robust control over mouse and keyboard events, while simultaneously enhancing the accuracy of gesture recognition [6].

In summary, the limb recognition system consists of these three pivotal modules, with each module employing specific algorithms or models. Their collaborative operation facilitates real-time human-computer interaction based on gestures.

2.2. Sensor-Based Gesture Recognition

A gesture recognition system, founded upon sensor technology and sEMG, comprises crucial components. Sensor selection, including cameras, depth cameras, inertial sensors, and sEMG sensors, is pivotal [7]. Data preprocessing enhances quality by mitigating noise. Core to recognition is feature

DOI: 10.54254/2755-2721/43/20230810

extraction, capturing attributes like hand position, motion, acceleration, and sEMG signal aspects. Machine learning (e.g., SVM, Random Forest) and deep learning (e.g., CNN, RNN) algorithms aid classification. sEMG preprocessing and feature extraction are specialized, involving filtering, time/frequency analysis, and signal fusion for informative cues. Personalized settings enhance adaptability. Interaction modules translate outcomes into actions. Real-time efficiency, user-friendly interfaces, security, and privacy are vital considerations. Integrating these modules forms a comprehensive gesture recognition system, broadening interaction applications.

2.3. Accuracy and Speed of Gesture Recognition

When reviewing gesture recognition studies, we identified two main problem categories. The first centers around addressing challenges when using standard equipment to capture specific, precise actions. Redundant channel-generated interference can complicate data processing and impair recognition accuracy. To tackle this, three solutions are proposed

Wei, W. et al. (2019) stated In the paper propose A multi-stream convolutional neural network (CNN) framework is proposed to learn the relationship between each specific action and the muscle movement required to form this action through a "divide and conquer" strategy to improve gesture recognition accuracy. Fig 1 show the hierarchical structure of a basic Convolutional Neural Network (CNN) [6].

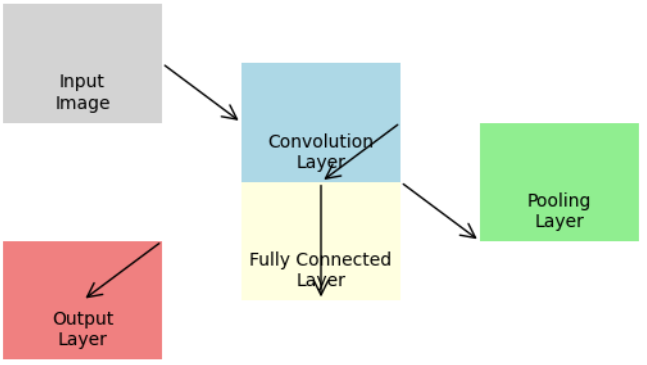

Figure 1. The hierarchical structure of a basic Convolutional Neural Network (CNN)

Human gesture recognition using surface electromyography (sEMG) is an important research topic in surface electromyography applications. Wei, W. et al. (2019) Using linear discriminant analysis (LDA) and extreme learning machine (ELM) to apply it to gesture recognition systems can improve the speed of processing complex signals and the accuracy of recognition by reducing redundant information in surface electromyographic signals [8].

Singha, J. et al. (2016) proposed a dynamic gesture recognition system for human-computer interaction through visual methods interaction [4]. The system consists of five stages: hand detection, hand tracking, feature extraction, feature selection, and classification. The hand detection technique is developed by combining three-frame differencing and skin filtering. Hands are then tracked using a modified Kanade-Lucas-Tomasi feature tracker, and by using different colors to mark and locate the final gesture area. Use ANOVA combined with incremental feature selection to select the best feature set and combine the individual classifiers (ANN, SVM, and kNN) to produce a classifier ensemble model.

The other type of question is how to accurately recognize human gestures or postures within the expected observation delay.Then for this kind of problem, we also found two solutions from the paper.

DOI: 10.54254/2755-2721/43/20230810

Zengeler, N. et al. (2018) proposed machine learning method based on deep data, using convolutional neural network and long-term short-term memory network to realize gesture recognition, providing new and more convenient control methods for driving assistance systems [7].

Sun, Y. et al. (2020) proposed that electrode redundancy in surface electromyography thumb movement recognition can be resolved using statistical variance theory [9]. More redundant channels can be removed while maintaining the same recognition accuracy. After removing channels, the training time of the classifier was reduced by nearly 20%.The Fig2 generated by this example code is a schematic diagram showing surface electromyography (sEMG) signals. SEMG signal is a bioelectrical signal generated by the electrical activity of muscles, which can be used to monitor the contraction and relaxation status of muscles.

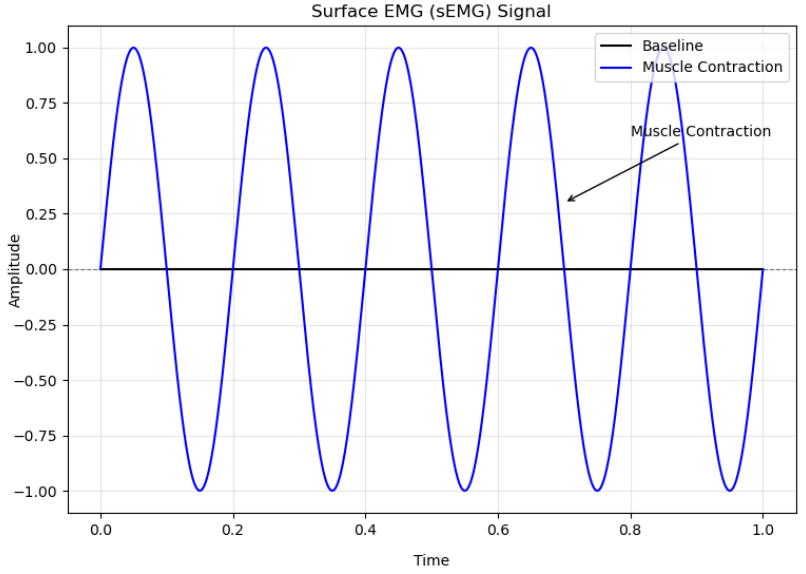

Figure 2. Surface Electromyography (sEMG) Signal Representation

2.4. The classical model/deep learning models

The research employs various deep learning algorithms, encompassing Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Residual Networks (ResNet).

CNNs are predominantly harnessed for the analysis of visual data, such as images or videos, adeptly learning and distilling salient features from raw input data. In gesture recognition, CNNs are leveraged to process gesture images and execute gesture classification tasks.

LSTM networks, a subtype of Recurrent Neural Networks (RNNs), are tailored for sequential data processing. They retain a memory of historical information, utilizing it to make predictions or decisions [10]. Renowned for handling extended dependencies and capturing patterns in sequential data, LSTM networks have found broad utility in domains like natural language processing, speech recognition, and time series analysis.

DOI: 10.54254/2755-2721/43/20230810

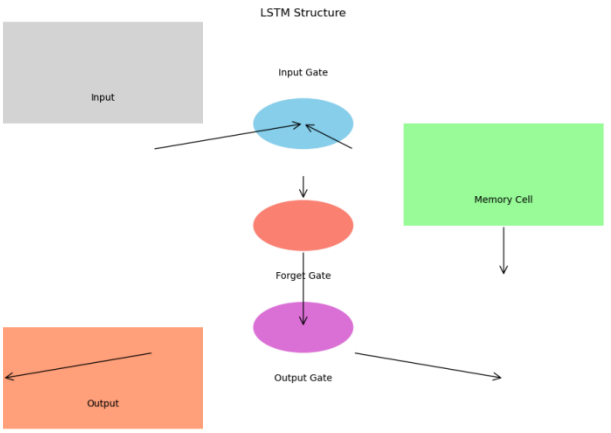

Figure 3. Illustration of LSTM Internal Structure

ResNet combines the residual and convolutional short-term memory models, creating a unified framework to extract spatiotemporal features globally and deeply. This fusion includes feature integration, preserving essential data. In our work, we utilize ResNet to enhance system accuracy.

Another significant approach involves the multi-stream residual network (MResLSTM), merging residual and convolutional short-term memory models. This amalgamation extracts spatiotemporal features comprehensively, incorporating feature integration to retain crucial data.

A multi-stream Convolutional Neural Network (CNN) framework, utilizing a "divide and conquer" strategy, emphasizes muscle-gesture correlation, heightening gesture recognition precision [2].

Furthermore, a vision-based approach is adopted, developing a dynamic gesture recognition system for human-computer interaction [4]. Comprising hand detection, tracking, feature extraction, selection, and classification stages, it integrates three-frame differencing and skin filtering for hand detection. Hand tracking employs adapted Kanade-Lucas-Tomasi feature tracking with color cues for accurate localization. Optimal features are determined via ANOVA, followed by fusion of individual classifiers (ANN, SVM, kNN) into an ensemble model.

2.5. The datasets used to develop the system

The data sets mentioned in these papers are basically image data sets. Most of these datasets are used to train intelligent gesture recognition systems so that they can make the most accurate recognition possible when faced with any recognized object.

The Type I dataset is used for the gesture recognition component, which is used to train the convolutional neural network (CNN), and it includes 3,200 different gesture image samples. The recognition system trained by the data set is capable of high-precision recognition. However, the database is too small to satisfy gesture recognition in more complex environments.

Type II data set is called REHAP data set, which consists of two unrelated parts: 600,000 sample images from 20 different people (REHAP-1) and 450,000 sample images from 15 different people (REHAP-2), 600,000 sample images from 20 different people (REHAP-1) and 450,000 sample images from 15 different people (REHAP-2), which will provide the basis for data-driven training. Improve speed and accuracy of gesture recognition. Compared with the previously mentioned data sets, it has a larger database and can provide the recognition system with more samples to help it accurately identify.

The third dataset type contains six dynamic gesture models aimed at improving the accuracy and stability of dynamic gesture recognition. Although the sample size is small, it has a more advanced sample type - the whole process of a person making a gesture, rather than a simple gesture image.

DOI: 10.54254/2755-2721/43/20230810

2.6. Application scenarios

This system is suitable for scenarios that require real-time high-precision gesture control, such as games, virtual reality, and human-computer interaction. It can recognize gestures through image processing and convolutional neural networks, and use Kalman filters to improve the stability and smoothness of the mouse cursor [10]. The application fields of this system mainly focus on virtual environments and interactive applications that require precise gesture control.

Furthermore, the proposed system holds promise across various domains, including computer gaming, sign-to-text translation, sign language communication, robotics, and video-based surveillance. Rigorous validation has been performed through one-way analysis of variance test, Friedman's test, and Kruskal-Wallis test. Notably, the proposed system exhibits favorable outcomes when compared to individual classifiers. The amalgamation of optimal features and classifier fusion culminates in the attainment of the highest overall accuracy.

At the same time, Intelligent human-machine interaction which was proposed in the paper based on electromyographic signals, a method for gesture recognition and human-machine interaction using surface electromyographic signals. It is mainly suitable for scenarios that require capturing information from muscle activity and achieving human-computer interaction. By recognizing electromyographic signals, especially thumb movements, intelligent human-machine interaction can be achieved, such as thumb gesture recognition. This method may be applicable to fields such as medical rehabilitation, intelligent assistive devices, and wearable technology to facilitate interaction between people with muscle damage or movement disorders and computer systems.

3. Analysis

Through careful reading and analysis of the above papers, we found that there are still some research gaps that have not been discovered and resolved, so we list them here so that we can conduct more in-depth research. These gaps can be roughly divided into two categories.

3.1. Model related research gap

First, how does the algorithm distinguish the real interaction intention of human beings from those accidental, inadvertent, and natural gestures? The phenomenon shown is that the user thinks that he has done the right gesture, but the system cannot recognize it correctly, or the user's inadvertent gesture is captured and executed by the system "accurately". There are many possible reasons behind it, such as the influence of environmental interference, the algorithm recognition threshold is too high, the recognition range is exceeded, the action is not standard, and so on. And when users need to pay time and attention costs unwillingly to smooth out these unexpected troubles, or can't get the expected correct response when they need to use it, then the interaction is really putting the cart before the horse.

Then, due to individual differences, such as gender, muscle state, and physical condition, it is difficult to obtain a general model for EMG pattern recognition. Therefore, it is often necessary to carry out long-term pre-training for individual users to obtain an accurate classification recognition model. Relatively time-consuming and labor-intensive. And in practical applications, the classifier generally does not change after the initial training, or remains unchanged for a long period of time. Moreover, sEMG also has time-varying characteristics, that is, as time goes by, Changes in the environment, changes in electrode positions and other factors will change the EMG characteristics, reducing the recognition accuracy of EMG signals.

3.2. Domain problem related research gap

So far, most of the research has focused on static gesture recognition technology, while dynamic recognition technology not only needs to track gestures, but also recognizes them. Its computational workload is heavy and slow, and it cannot be used in real-time recognition systems. Moreover, complex background interference, difficulty in identifying and locating fingertip feature points, and changes in the lighting environment will all lead to a reduction in the recognition rate of dynamic gesture

DOI: 10.54254/2755-2721/43/20230810

recognition. Moreover, the expression trajectories of some dynamic gestures based on monocular vision are similar, and it is easy to have certain misjudgments.

Besides, the expression of gesture meaning has been influenced by different cultures. Seemingly simple gestures contain countless possibilities, and people of different ages have different perceptions of the same action. Different gestures have completely different meanings in different countries. For these differences, the solution given in the existing articles is only to distinguish through the settings before each use, and does not give a good solution for the use of a system by different people.

At the same time, in the gaming and entertainment industry, gesture recognition offers immersive experiences. Addressing challenges related to latency, seamless integration, and accurate recognition of dynamic gestures will enhance the realism and engagement of these applications.

4. Future direction

For these gaps, we consider that in the realm of sensor-based gesture recognition, a noticeable research gap lies in addressing the challenges posed by real-world variations, environmental dynamics, and user diversity. The current advancements in sensor-based gesture recognition systems have indeed made remarkable strides; however, several key challenges remain to be effectively tackled:

Adaptation to Real-World Variations: The performance of sensor-based gesture recognition systems often encounters difficulties when dealing with real-world variations such as lighting conditions, background clutter, and user-specific nuances. The existing research has yet to comprehensively address the robustness of these systems across diverse scenarios and dynamic environments, necessitating the development of adaptive algorithms that can consistently deliver accurate results despite varying conditions [6].

Dynamic Environmental Changes: Real-world environments are dynamic and can introduce unexpected changes that impact gesture recognition accuracy. The current systems often lack the flexibility to adapt in real-time to dynamic environmental changes, leading to diminished recognition performance. Investigating methods to enhance the adaptability of these systems to changing conditions, such as leveraging reinforcement learning or adaptive algorithms, remains a promising avenue for future research.

User Diversity and Personalization: User diversity introduces challenges in terms of variations in user characteristics, body shapes, gestures, and preferences. The existing approaches often struggle to achieve consistent performance across diverse user profiles. Developing personalized gesture recognition models that can adapt and cater to individual users' unique characteristics could bridge this gap, leading to more accurate and adaptable recognition systems.

Unconstrained Gesture Recognition: Many real-life scenarios involve unconstrained gestures that are not predefined or limited by a fixed set of actions. Existing gesture recognition systems may lack the flexibility to recognize these spontaneous and diverse gestures accurately. Addressing the gap of unconstrained gesture recognition requires innovative approaches that can generalize well to different and potentially unexplored gestures.

In conclusion, the research gap in sensor-based gesture recognition revolves around effectively adapting to real-world variations, addressing dynamic environmental changes, catering to user diversity, and tackling unconstrained gesture recognition challenges. By overcoming these challenges, the field can advance towards creating more robust, versatile, and user-centric gesture recognition systems that seamlessly integrate with various applications and environments.

5. Conclusion

Gesture recognition has significant applications in various fields, including computer vision-based limb recognition systems and sensor-based gesture recognition systems. These systems enable real-time high-precision gesture control, making them suitable for applications such as games, virtual reality, human-computer interaction, sign-to-text translation systems, sign language communication, robotics, and video-based surveillance.

DOI: 10.54254/2755-2721/43/20230810

Computer vision-based gesture recognition systems rely on hand detection, gesture recognition, and recognition-driven human-computer interaction modules. These systems utilize algorithms such as Convolutional Neural Networks (CNN) for gesture recognition and mapping tables to translate recognized gestures into specific mouse or keyboard events.

Sensor-based gesture recognition systems use various sensors, including cameras, depth cameras, inertial sensors, and surface electromyography (sEMG) sensors, to capture accurate gesture data. These systems involve data collection, preprocessing, feature extraction, classification, calibration, personalized settings, and interaction modules. Machine learning algorithms (e.g., SVM, Random Forest) and deep learning techniques (e.g., CNN, RNN) are applied for gesture classification.

But there are some research gaps in this area. Such as distinguishing real interaction intentions from accidental, inadvertent, and natural gestures, and addressing individual differences in EMG pattern recognition.

To address these research gaps, we suggested to develop adaptive algorithms that can handle real-world variations and dynamic environmental changes and investigate methods for personalized gesture recognition models to cater to individual user characteristics. Whatsmore, explore innovative approaches for unconstrained gesture recognition and enhance the adaptability and robustness of gesture recognition systems across diverse scenarios and user profiles can be helpful.

In conclusion, gesture recognition technology has made significant advancements, but there are still challenges to overcome. By addressing the identified research gaps, the field can advance towards creating more robust, versatile, and user-centric gesture recognition systems that seamlessly integrate with various applications and environments.

Acknowledgement

Shijia Li, Luoyun Zhou, Mingqi Fan, and Yucheng Xiong contributed equally to this work and should be considered co-first authors.

References

[1] Wei, W. et al. (2019). A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface.

[2] Xu, W. (2017). Research on Gesture Recognition in Human-Computer Interaction Based on Convolutional Neural Network. Proceedings of the 2017 2nd International Conference on Image, Vision and Computing (ICIVC).

[3] Yang, Z. et al. (2021). Dynamic Gesture Recognition Using Surface EMG Signals Based on Multi-Stream Residual Network.

[4] Singha, J. et al. (2016). Dynamic hand gesture recognition using vision-based approach for human–computer interaction.

[5] Xu, P. (2017). A Real-time Hand Gesture Recognition and Human-Computer Interaction System.

[6] Hazra, S., & Santra, A. (2018). Robust Gesture Recognition Using Millimetric-Wave Radar System.

[7] Zengeler, N. et al. (2018). Hand Gesture Recognition in Automotive Human–Machine Interaction Using Depth Cameras.

[8] Qi, j. et al. (2019). Intelligent Human-Computer Interaction Based on Surface EMG Gesture Recognition.

[9] Sun, Y. et al. (2020). Intelligent human computer interaction based on non-redundant EMG signal.

[10] D. Jiang, G. Li, Y. Sun, J. Kong, B. Tao, and D. Chen, “Grip strength forecast and rehabilitative guidance based on adaptive neural fuzzy inference system using sEMG,” Personal and Ubiquitous Computing, 2019.