Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/77/20240487

Implementation and research of gesture recognition in HCI field

Haoran Zhou[1,3,*], **Yixiang Ren**[2,4]

[1]Department of Information Science and Engineering, Wuchang Shouyi College, Wuhan, 430064, China
[2]Department of Information Engineering, Hangzhou Dianzi University, Hangzhou, 311305, China

[3]1643276687@qq.com
[4]renyx0904@gmail.com
*corresponding author

Abstract. Since the first Neural Network computer was made by Marvin Minsky and his schoolmates in 1951, the development of artificial intelligence (AI)has undergone various and huge changes. It generally evolves into a field that has infinite possibilities and a major branch of computer science that can not be ignored. With the trend of computerization and the rapid evolution of the Internet of Things (IoT), gesture recognition emerged. Various prototypes are blossoming in the laboratories, and some of them have become certain products that have practical applications in later days. At the same time, more and more deep learning technology has been applied to gesture recognition systems that greatly improve the quality of the service. The purpose of this review is to summarize and analyze the existing algorithms of Dynamic Gesture Recognition systems, which have several different methods that are based on multiple signal extractors. This study focuses on the 3D Convolutional Neural Network(CNN), which plays an important role in the Dynamic Gesture Recognition optimization algorithm and data analysis algorithm. In this paper, we reviewed past papers in the gesture recognition field, which include the sEMG method, microwave method, and vision recognition method. During the process of data collecting and summarizing the past papers, we mainly focused on the accuracy differences between recognition methods and the efficiency differences in data processing algorithms. (CNN-based, LSTM based, etc.) Then, we analyzed the main difficulties and challenges of these methods, which are briefly listed in the Introduction part. Data processing algorithms are also being studied, and a horizontal comparison between CNN-based, LSTM-based, and transformer-based algorithms is also being made. Besides this, for those problems that already have a reliable solution, we also summarized the possible solutions and listed them out. We found that the gesture recognition system has already been systematically studied and is partly used in some fields. However, the algorithm and modeling methods can still be optimized and it also needs further study to be more widely used.

Keywords: human-computer interaction, gesture recognition, CNN, LSTM, Transformer, sEMG

1. Introduction

As an effective way to communicate between humans and computers nonverbally, gesture recognition is one of the most studied Human-Computer interaction(HCI) methods in recent years, which contains multiple branches; various methods are derived from these branches that are applicable in different work environments. For our research on hand gesture recognition systems, we summarized various approaches in the past few years; the range is covered from mainstream methods like surface electromyography(sEMG) gesture recognition and vision-based gesture recognition to relatively less popular methods like millimeter-wave radar-based gesture recognition.

Prominent examples of the gesture recognition field is dynamic hand gesture recognition, which is mainly based on the sEMG method and vision method, so this review mainly focuses on these two methods and summarizes the past literature in the gesture recognition field.

With the progress of computerization and the development of the Internet of Things (IoT), the traditional input/interact methods such as keyboard and touch screen are incompetent under complex application scenarios like controlling multi-devices at the same time without physical contact or individual controllers. Besides this, the physical input methods can not meet the requirement of efficiently interacting with modern multifunctional AI, whose information exchange density is extremely high. A novel control/interact method is needed to deal with the complex interaction.

The structure of the paper is as follows:

1. First, we identify and summarize the key components of every gesture recognition system.
2. Then, we analyze different algorithms which structure different methods.
3. Then, we systematically compare different datasets for training the system. The datasets mainly come from the OpenCV and NinaPro databases.
4. Finally, we identify some research gaps that provide insight for future research direction.

However, limited by time and our personal capability, there are several unavoidable shortages on the choice and analysis of the literature, which mainly shown in the study of databases

2. Literature Review

2.1. Significant Step

Human-Computer Interaction (HCI) research, including gesture recognition, proceeds through a series of stages that collectively drive the investigation and advancement of gesture-based interaction paradigms. We conducted a detailed literature analysis after identifying the characteristics of the research topic. Then, we gathered insights into previous stages of development in the field, future trends, and the gap between research and practical applications.

The following stage involves data collecting and preparation, which includes the use of various sensors or devices to record pertinent gesture data. This comprises motion capture systems, depth cameras, gloves, and wearable accelerometer-equipped devices. To ensure data quality and dependability for later studies, the raw data is categorized, structured, cleaned, and annotated.

The phases of feature extraction and representation, which transform raw gesture data into computationally detectable features, become critical. Coordinates of key locations, joint angles, motion trajectories, and electromyographic (EMG) signals. For example, it can be retrieved and quantified from a gesture image. EMG data was collected by inserting electrodes on muscles. It can record muscle electrical activity and can be used to examine muscle contractions, relaxation, and movements. These characteristics serve as the foundation for additional study and model building.

Algorithm selection and implementation are both critical. Researchers select algorithms that are relevant to their study goals and employ appropriate gesture recognition techniques. Commonly used machine learning methods include support vector machines, decision trees, and deep learning architectures such as convolutional neural networks.

Following that, the procedure moves on to model training and evaluation. A subset of the dataset is set aside for training, while the remainder is used to evaluate the model's generalization capability. To assess the model's correctness and efficacy in accurately identifying gestures, performance metrics such as accuracy, precision, recall, and overfitting are used.

The final stage is to do simulation experiments. These studies imitate real-world circumstances in order to assess the effectiveness of the recognition system. Participants engage with the system by making real-life motions. Then, the results provide feedback on the model's strengths, flaws, and places for improvement.

2.2. Difficulties Review

2.2.1. Difficulty 1

For sEMG recognition methods, the main issue is the negative effect on recognition accuracy caused by time non-stationarity and redundancy of sEMG.

For example, the accuracy of recognition will be greatly reduced under intermittent usage, which is mainly because of the time variability during signal acquisition progress. During the movement progress of the forearm muscle group, the electromyographic signals are discontinuous; this will cause a time difference in the signal that will be acquired. Besides this, time's non-stationarity can also generate redundant interface signals, which will lead to the redundancy of sEMG time domain information and finally lead to a decline in recognition accuracy.

Another factor of redundancy is mainly related to acquisition equipment. For equipment composed of 16 electrodes, significant redundancy can still be observed in the part outside the blue area [1].

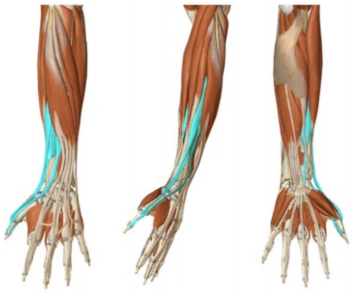

Figure 1. Surface EMG signal generation region

Solution: Jinxian Qi et al. came up with a dimensionality reduction method based on Linear discriminant analysis (LDA) in 2019**.** LDA and extreme learning machine (ELM) are used to reduce the dimensionality of high-dimensional signals and eliminate the redundant information in sEMG [2].

To reduce the dimensionality, two new dataset are needed, sEMG of nine static gestures which collected in two days are averaged to get the new dataset X2. The process of feature fusion can be represented as follow equation:

$$X2 = [RMS1 + RMS2, WL1 + WL2, MAS1 + MAS2]$$

Another dataset X3 will be applied to dimension reduction, which is obtained by normalizing X2. LDA is used to eliminate the matching problem between channels and features, while weakening some channels with less correlation or more redundant information and improving the recognition rate. The process of dimension reduction is as follows.

Input dataset X2 = [x1, x2, . . . , xc], where any sample xi is a c-dimensional vector. Computing the intra-class divergence matrix Sω. Computing inter-class divergence matrix Sb Computational matrix S

Calculating the maximum d eigenvalues of S−1ω Sb and cor- responding d eigenvectors ω1, ω2, . . . , ωd and the projection matrix W. For each sample xi dataset, it is transformed into a new sample zi = WTxi.

Output sample set X3 = [z1, z2, . . . , zc].

After dimension reduction by LDA, the dimension of fused features becomes d [3].

With the above process, the high-dimensionality signal with c-dimensions is changed into d-dimensions. The signal after the dimension reduction process can greatly optimize the data processing algorithm.

2.2.2. Difficulty 2

Although a relatively high accuracy rate can be reached on static gesture recognition, it still can not meet the requirements of human-computer interaction because, in the actual usage scenarios, there's almost no way to find a static gesture that waits to be recognized. Besides this, for daily users, the recognition system must be fast and won't be inconvenient. In other words, what people need is a stable and efficient real-time dynamic gesture recognition system. That brings two difficulties: Real-time recognition and the limit of hardware.

For the first difficulty, of course, performance sensors can be one of the solutions, but the problem is that these sensors perform well but are costly, which might not be suitable for personal use [2].

Solution: According to Pei Xu et al., their team put forward a real-time gesture recognition and HCI method based on CNN, which is mainly composed of a CNN filter combined with an image processing step to recognize gestures.[2] Besides, Xu's team introduced the Kalman filter into the mouse cursor control process to smooth the motion of the mouse cursor controlled by hand gestures. Different from the sEMG method, Xu's team used a vision-based method to implement gesture recognition, which uses a series of sequential, static 2D images provided by one monocular camera. During this progress, they performed background removal on the hand, turning it into black and white and polygon to get clearer shape and color features.

2.3. Classical Model

2.3.1. CNN Based

CNN (Neural Convolutional Networks) is a deep learning model used to analyze photos, video, and other visual data. It is one of the most widely used models in computer vision, picture identification, target detection, picture segmentation, and other computer vision tasks that make extensive use of it. The CNN algorithm has many essential components, including a Convolution Layer, Pooling Layer, Activation Function, Full Connection Layer, and Weight Sharing. In practice, input data is convolved to build a sequence of feature graphs. It uses a set of learnable filters (also known as a convolution kernel). Different features in the input data can be detected by each filter. The output of the convolutional layer can be used as the input to the next layer. Then the network can gradually extract more abstract and high-level characteristics by stacking numerous convolutional layers. The pooling layer then reduces the resolution of the feature map, which is output by the fully connected layer. CNN, by reducing the amount of parameters in the network through local connections and weight sharing. It can also effectively lower the occurrence of overfitting and improve the model's generalization capacity. The pooling layer compresses the image resolution minimizes the amount of data processed and increases the model's computational efficiency.

In their study, Wei et al. applied deep learning algorithms for gesture-based sEMG recognition and suggested a new multi-stream CNN framework to increase accuracy [4]. In the experiment, Wei evaluated three different sEMG databases: the NinaPro database, the CSL-HDEMG database, and the CapgMyo database [4]. Wei employs the MxNet framework to create multi-stream CNNS [4]. It is trained using stochastic gradient descent (SGD). During the multi-stream decomposition phase, the original sEMG image is divided into equal-sized chunks (streams). Then, each of these chunks independently learns representative features using CNNS. The features learned from all of the streams fuse into a uniform feature map. After that, it will subsequently sent into the fusion network for gesture detection. Wei used the same pre-training method as in the prior study during the training phase [4]. The CNN model is pre-trained on each cycle of training data. Then the pre-training parameters are utilized to initialize the parameters of the multi-stream CNN. Wei tested the performance of CNN using multiple variants of the input sEMG image patch and varied time windows and majority voting windows in the experiment [4]. Wei discovered that the multi-stream CNN framework demonstrated great accuracy in sEMG gesture-based recognition by comparing experimental data [4]. The system delivers outstanding performance in gesture detection tests on several databases, proving the framework's capacity to generalize across data sets. Wei also compares the framework to other CNN-based approaches, such as deep domain adaptation and self-calibration categorization [4]. As a result, the multi-stream CNN system offers considerable benefits in terms of accuracy and resilience. The deep learning method is effectively applied to the sEMG gesture-based recognition task utilizing the multi-stream CNN framework. And finally produced satisfactory results.

They used DCGans (Deep Convolutional Generative Adversarial Networks) to create vast volumes of gesture image data in the Fang experiment [5]. This data was used to train the CNN model. They employed image data for 37 gesture categories during training, totaling 44,400 photos [5]. Of these, 37,000 were used for training and the remaining 7,400 were used for testing [5]. Fang employs SGD (Random Gradient Descent) as an optimizer and sets default parameters to maximize training efficiency [5]. There were also twenty epochs and a batch size of 500 [5]. Fang enhances gesture recognition accuracy by increasing the number of layers and depth of CNNS [5]. According to test results, cutting the number of frames from 30 to 20 boosted recognition speed by nearly a factor of 1, but only 6% of accuracy was lost [5]. The recognition speed increased by almost two times when the number of frames was cut from 30 to 10, although the accuracy only dropped by around 10% [5]. This means that their model can recognize and produce results practically instantly. It can also address the needs of people's daily lives. In order to further confirm the accuracy of the test results, Fang also examined individuals who did not take part in the sampling [5]. The experimental findings show that the CNN-based gesture detection model performed well in real-time and under a variety of lighting scenarios [5]. In this study, experiments utilizing the CNN model are used to demonstrate its efficacy and viability in gesture recognition [5].

A convolutional neural network (CNN) that was developed from LeNet-5 was employed in Xu's article to recognize motions [2]. The CNN classifier has a straightforward structure. This structure consists of two convolutional layers, two fully linked layers, and a ReLU activation function. The features of the CNN classifier learns using the maximum pooling layer exhibit specific rotation invariance. In order to prevent the classifier from being influenced by the color attributes of the hand, the CNN is fed a binary image as input. The image must be preprocessed before being sent to the CNN classifier. The pre-processing steps include thresholding, morphological transformation (open-close operation), background subtraction (optional), hand color filtering, Gaussian blur, and contour extraction. After pretreatment, the hand detector began to work. It tracks the movement of a particular point on the hand. After that, it estimates the center and palm radius of the hand. Then, it extracts the contoured area of the hand. Finally, adjust the outline area to a fixed size and maintain the aspect ratio. Then, center on the canvas and enter the CNN classifier. The center of the hand is estimated by distance transformation. After the distance transformation, the pixel point with the maximum value in the hand contour region is considered to be the center of the hand. The distance between the hand's center and the effective convex defect's farthest point is used to estimate the palm radius. Xu employed gestures to use the mouse and keyboard and the ROS robot operating system to control the walking of a virtual robot in terms of human-computer interaction. Xu also utilizes a Kalman filter to smooth the mouse's movement in order to increase the accuracy of the mouse control [2]. In general, Xu uses a convolutional neural network (CNN) to recognize gestures because it can do so with a high degree of accuracy and in real-time [2].

2.3.2. LSTM Based

LSTM stands for long short memory term network. It is a type of recurrent neural network (CNN). The standard RNN model suffers from gradient disappearance, which means that the gradient gradually decreases until it is too small to contribute to neural network learning. And because this occurs at earlier levels, RNNs are unable to learn these levels well. It will discard the information and retain just short-term memory. The LSTM's peculiar cell state and multiple gates effectively handle the problem of short-term memory. The gating unit is separated into four sections: the forgetting gate, the input gate, the update memory state, and the output gate. LSTM enables the network to selectively recall, forget, and output information via these gating processes. So it can allow it to handle long-term dependencies more efficiently. As a result, it excels in numerous sequence modeling applications. Data such as finger and palm pressure, Electromyography (EMG), and inertial measurement Unit (IMU) can be gathered and translated into a format suited for LSTM, such as a time-step X feature matrix. Then, a suitable LSTM model is used, and the data is utilized to train the model to recognize various gestures. The model can extract gesture models and features from time series data, increasing the accuracy and robustness of gesture detection technologies.

A gesture recognition system based on inertial sensors and deep learning technologies was introduced by Mali et al. [6]. They discussed the value of gesture recognition in human-computer interaction and outlined a few issues with the state-of-the-art gesture recognition techniques. Additionally, they unveiled an end-to-end gesture recognition system. This system uses to deep learning and transfer learning models to recognize movements in real time [6]. For control reasons, gestures are wirelessly sent through Bluetooth to a personal computer [6]. The system is genuinely customizable because users can combine their own gestures with a library of pre-existing gestures [6]. They developed an all-purpose desktop program called the Action Mapping Interface to test the system [6]. It replicates keyboard and mouse actions and executes complicated commands using gestures picked up by neural networks [6]. They developed an effective gesture recognition model using the LSTM algorithm [6]. In order to categorize various motions in the model, they collected data from electromyogram (EMG), inertial measurement unit (IMU), and finger and palm pressure data [6]. Additionally, Mali et al. created a DIY armband to gather IMU data and feed it into the LSTM model [6]. They also suggest a wearable device based on a MYO armband for continuous gesture detection and a programmable array of pressure sensors [6]. They can more precisely acquire things in the user's hands because of this array of pressure sensors. The results of the experiment indicate that using the LSTM model to recognize gestures can increase accuracy.

The classification of medical gestures using LSTM recurrent neural networks is the main topic of the paper by Cifuentes et al. [7]. In order to begin the trial phase, the researchers first gathered data collection from 14 obstetricians, including eight new doctors and six junior physicians [7]. A total of 392 gesture locus samples were provided by each doctor who placed 6–20 forceps blades [7]. To categorize these motions, they employed the LSTM (Long Short-term Memory) algorithm. The first step in preprocessing gesture data is normalization and smoothing. Following that, they divided the data set into a training set and a test set, using 70% of the data for training and 30% for testing [7]. Using a training set, the researchers created an LSTM model and trained it [7]. To increase classification accuracy, the LSTM network gradually modifies its internal parameters during training. Finally, they used a test set to assess the trained LSTM model [7]. They determined the classification accuracy of the model [7]. It is expressed as the proportion of gesture samples that were properly classified to all samples. The experimental findings led Cifuentes et al. to the conclusion that the LSTM model is quite accurate in classifying medical gestures, achieving a classification rate of 99.1% [7]. The model offers advantages in how it approaches time series issues and teaches long-term dependence, and it has a wide range of possible applications in the field of medical education.

2.3.3. Transformer Based

Vaswani et al. presented Transformer, a neural network architecture for processing sequence data, in 2017 [8]. Used to solve tasks that need sequence-to-sequence matching, such as machine translation, language development, text summarization, and so on. Transformer differs from typical recurrent neural networks (RNNS) and recurrent units (such as LSTM and GRU) in two important concepts: Self-Attention and Multi-Head Attention. Transformer also includes a new acyclic structure that does not rely on explicit loop connections but rather on a self-attention method to record relationships between sequence positions. In order to better capture dependencies over vast distances, this attention technique can establish a connection between any two positions in the sequence. The relevant skeletal data, image sequences, or video resources may be captured and input the model. The transformer model can capture and learn the time-series properties and context of movements in the field of gesture recognition. If the input data is multimodal, the transformer can handle several forms of input data and learn these properties in a single framework. Traditional sEMG signal recognition approaches often employ convolutional neural networks (CNN) to extract features; however, CNN suffers from local perception bias and is incapable of capturing global context information. To achieve better outcomes, employ the Transformer model to extract crucial channel information via the attention mechanism.

A sEMG signal identification technique based on the channel-level transformer model is introduced in this study by Matsuda et al. [9]. Convolutional neural networks (CNNs) are frequently used in traditional sEMG signal recognition techniques to extract features, but CNNs suffer from local perception bias and are unable to capture global context data. Through the attention mechanism, the method put forth by Matsuda et al. uses the Transformer model to retrieve crucial channel information [9]. Data sets and preparation are done first in the experimental phase. They used the Ninapro benchmark dataset (DB2) [9]. 40 healthy volunteers collected the data, recording sEMG signals in 12 channels at a sample rate of 2 kHz in order to find 17 gestures [9]. Each subject's signal was subjected to a Butterworth bandpass filter (5-500 kHz, order 3) [9]. Then, it was used for segmentation and sample creation with a 200-millisecond moving window, using a total of 81,732 sEMG samples [9]. The extraction of time-frequency features is then performed [9]. Each raw data set was transformed into a time-frequency domain signal using a two-tree complex transformation, and each sample's feature matrix was created using six traditional sEMG feature extraction techniques.[9] The following outcomes are attained in the end: The method provided by Matsuda et al. has a very high accuracy rate of 85.6% in gesture recognition tasks by employing the Transformer model to model channel information [9]. The authors' method outperforms conventional CNN models, even when only six classical features are recovered without the use of intricate CNN filters. The suggested method outperforms the conventional CNN method in gesture recognition tasks, according to experimental findings. This demonstrates the Transformer model's potential for sEMG signal identification.

Le et al. [10] studied the performance of the Transformer model for motion identification on inertial sensor data as well as the usage of inertial sensors for the detection of human motion. For their analysis, they employed the three publicly accessible datasets: CMDFALL, C-MHAD, and DaLiAc.[10] Le et al. pre-processed the raw data through a data loader and fed it into the framework during the pilot phase [10]. This pre-processing included eliminating any duplicated or missing data gestures and labeling the data. Then, the Transformer model is used to extract potential embeddings from the raw data, and multi-layer perceptrons (MLPS) are used to categorize them [10]. Pre-processing, Transformer coding, and classification make up the framework's three phases. The performance of the chosen varied window sizes when applied to the Transformer model was compared to the use of various window sizes (w = 64, 96, and 128) on each dataset [10]. To support the effectiveness of the Transformer model, they also compared the findings of earlier investigations [10]. The outcomes demonstrate that the performance of the model is affected in certain ways by various window sizes [10]. Choosing the proper window size can dramatically increase recognition accuracy in CMDFALL data sets. The Transformer model increased F1 scores in the S3 group of the CMDFALL dataset by 19.04% above conventional techniques [10]. The Transformer model's accuracy in the C-MHAD dataset was 99.56% [10]. On these three datasets, you can observe that the Transformer model produces better outcomes.

2.4. Different Application Field

Gesture recognition technology has advanced dramatically in recent years. It enables users to communicate with computers and devices using natural gestures and movements, eliminating the need for physical contact or external devices like keyboards and mice. This improves the user interface's intuitiveness and usability. Scholars predicted that gesture recognition technology would be useful in a wide range of industries.

This development of gesture recognition technology in the realm of medicine has enhanced medical diagnosis, rehabilitation treatment, surgical operations, and the operation of medical equipment. Gesture recognition technology, for example, can be utilized to navigate and support procedures in the operating room. Surgical navigation technology can assist surgeons in performing more precise operations during surgery. Surgeons can utilize motions to alter images on a computer screen, zoom in and out of the images, and rotate the viewing Angle. So they can better grasp the patient's anatomy and guide surgery. Assisted surgery is a robot-assisted surgical approach in which the surgeon's orders are executed by a robotic system to improve surgical accuracy and stability. Gestures, such as moving the robotic arm to a specified location or spinning the instrument, can be used by the surgeon to direct the movement of the surgical robot. Some patients can interact with rehabilitation equipment via gestures to exercises. The virtual operation and training of medical models by medical staff via gesture recognition. There are successful applications of gesture recognition technology in the medical field.

In the realm of education, for example, the advent of gesture recognition technology has changed the interaction of the educational process. It can increase students' participation in class so that they obtain better teaching results. Gestures can be used by students to respond to questions, participate in group discussions, and interact with educational programs. Gesture recognition can also benefit children with special educational needs. And it can help them interact with educational content more effectively. People with autism can utilize gestures to express their needs and emotions. At the same time, gesture recognition can be integrated with VR technology to improve educational results. Students can use VR devices to visit a virtual laboratory. All teachers and students can experiment with gestures without the need for real laboratory equipment. This can not only successfully alleviate the problem of inadequate educational equipment but also allow every student to have a thorough understanding of the experimental equipment. It will not bring about the repercussions of the experiment when it goes wrong, not only improving safety but also giving the kids a very intuitive grasp of the dangerous component of the experiment.

3. Analysis

A significant problem in machine learning is overfitting, which occurs when a model performs well on training data but badly on fresh, untainted data. Overfitting has the following causes:

Model Complexity: A model that is overly complicated may perform poorly on new data. Then, it will be able to "memorize" every nuance in the training data. Having too many parameters, precisely fitting the training set, and disregarding generalization are all causes of excessive complexity.

Insufficient Training Data: The model may struggle to generalize to new data due to limited data samples that don't reflect the entire data distribution.

Capture of Noise and Irrelevant Details: The model tries to adapt to every data point, including noise or outliers. Because its goal is to decrease errors on the training data. As a result, the model overfits to the data fluctuations by incorporating noise and unimportant characteristics. This leads to false predictions.

Overfitting is significant because it has the potential to weaken a model's applicability in the actual world. While an overfitted model may achieve high accuracy on the training set, it may struggle to generate correct predictions on new data, limiting its reliability and applicability in real-world applications.

Overcoming the problem of overfitting is quite difficult. The difficulty of feature engineering restricts the model's comprehension of the data. It requires more work in feature extraction and data preprocessing. Data cleaning and preprocessing become essential due to the flaws in real-world data. So, careful handling of noise and outliers is necessary.

The neural network-based recognition technique has a high ability to categorize and identify, according to Fang et al. [5]. However, if the number of neural network layers is generally shallow, overfitting is easy. To address the overfitting problem, they developed a gesture recognition approach based on CNN and DCGAN. Overfitting difficulties caused by photos in the dataset being too similar can be prevented by using DCGAN to provide different training data. The experimental results show that using the CNN+DCGAN model for gesture recognition can significantly improve the accuracy [5].

4. Future Direction

I think that as human-computer interaction becomes more prevalent, gesture recognition technology will advance. More precision and dependability will be seen in gesture recognition technologies in the future. System comprehension of user gestures will be enhanced by the development of machine learning and deep learning algorithms, which will increase the accuracy of gestures. Future gesture recognition technologies will also be more dependable and accurate. System comprehension and interpretation of user gestures will be enhanced by the development of machine learning and deep learning algorithms, which will increase the accuracy of gestures. Perhaps in the future, more sophisticated sensing technologies—like the ability to detect minute muscle movements and the lately well-liked brain-computer interface—will enhance the sensitivity and variety of gesture interactions even more. Of course, privacy and security concerns will also be put to the test. Because technology is developing so quickly, it is also essential to protect user data well in order to avoid privacy issues. Future gesture interaction technologies will, in general, be more widely used and diverse, offering users a more organic, simple, and customized means of communication. This will spur technological advancements and improvements in human-computer interaction, among other areas.

5. Conclusion

To sum up, through the study of past papers, which mainly focused on the methods and algorithms of the gesture recognition system, we got a lot of precious first-hand information. Through these past papers, we think the sEMG extractor is a significant part of the signal input end; compared with microwave radar or vision image recognition methods, it has the advantages of accuracy and robustness. And it has the potential to become a mainstream gesture recognition method in the future [11,12]. However, it still has difficulties in daily practical applications due to it needing device support to work, which might cause inconvenience in daily use. Fortunately, the miniaturization of wearable smart devices like iWatch, which has been equipped with sEMG extract function in recent years, shows that technological progress is gradually reducing this shortage.

For the algorithm, we compared and summarized the following Neural Network models:

(1). CNN based model

(2). LSTM based model

(3). Transform based model

According to papers related to these different models, CNN models are more focused on sEMG-based gesture recognition and have the ability to realize high-accuracy dynamic gesture recognition. The LSTM model is good at medical gesture recognition, there are already several teams studying in this field. The Transform model is one of the models that can erase the local perception bias issue and the inability to capture global context data that CNN suffers.

References

[1] D. Kaneishi, R. P. Matthew and M. Tomizuka, "A sEMG Classification Framework with Less Training Data," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 1680-1684, doi: 10.1109/EMBC.2018.8512623.

[2] Pei Xu, (2017), A Real-time Hand Gesture Recognition and Human-Computer Interaction System. [online]Available at https://arxiv.org/abs/1704.07296 (2/8/2024 22:00). J. Qi, G. Jiang, G. Li, Y. Sun and B. Tao, "Intelligent Human-Computer Interaction Based on Surface EMG Gesture Recognition," in IEEE Access, vol. 7, pp. 61378-61387, 2019, doi: 10.1109/ACCESS.2019.2914728.

[3] J. Qi, G. Jiang, G. Li, Y. Sun and B. Tao, "Intelligent Human-Computer Interaction Based on Surface EMG Gesture Recognition," in IEEE Access, vol. 7, pp. 61378-61387, 2019, doi: 10.1109/ACCESS.2019.2914728.

[4] Wentao Wei, Yongkang Wong, Yu Du, Yu Hu, Mohan Kankanhalli, Weidong Geng,
A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-

computer interface, Pattern Recognition Letters, Volume 119,2019, Pages 131-138, ISSN 0167-8655,https://doi.org/10.1016/j.patrec.2017.12.005.

[5] W. Fang, Y. Ding, F. Zhang and J. Sheng, "Gesture Recognition Based on CNN and DCGAN for Calculation and Text Output," in IEEE Access, vol. 7, pp. 28230-28237, 2019, doi: 10.1109/ACCESS.2019.2901930.

[6] D. Mali, A. Kamble, S. Gogate and J. Sisodia, "Hand Gestures Recognition using Inertial Sensors Through Deep Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579829.

[7] J. Cifuentes, P. Boulanger, M. T. Pham, F. Prieto and R. Moreau, "Gesture Classification Using LSTM Recurrent Neural Networks," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 6864-6867, doi: 10.1109/EMBC.2019.8857592.

[8] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, \L ukasz and Polosukhin, Illia, Advances in Neural Information Processing Systems, Curran Associates, Inc., Attention is All you Need,https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf, volume 30, 2017.

[9] J. Zhang, Y. Matsuda, M. Fujimoto, H. Suwa and K. Yasumoto, "Feasibility Analysis of sEMG Recognition via Channel-Wise Transformer," 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2022, pp. 105-106, doi: 10.1109/GCCE56475.2022.10014071.

[10] T. -H. Le, T. -H. Tran and C. Pham, "Human action recognition from inertial sensors with Transformer," 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Phu Quoc, Vietnam, 2022, pp. 1-6, doi: 10.1109/MAPR56351.2022.9924794.

[11] Ying Sun, Chao Xu, Gongfa Li, Wanfen Xu, Jianyi Kong, Du Jiang, Bo Tao, Disi Chen, Intelligent human computer interaction based on non redundant EMG signal, Alexandria Engineering Journal, Volume 59, Issue 3,2020, Pages 1149-1157, ISSN 1110-0168,https://doi.org/10.1016/j.aej.2020.01.015.

[12] Yang Zhiwen, Jiang Du, Sun Ying, Tao Bo, Tong Xiliang, Jiang Guozhang, Xu Manman, Yun J untong, Liu Ying, Chen Baojia, Kong Jianyi, Dynamic Gesture Recognition Using Surface E MG Signals Based on Multi-Stream Residual Network, Frontiers in Bioengineering and Biot echnology, VOLUME9, 2021, https://www.frontiersin.org/articles/10.3389/fbioe.2021.779353, 10.3389/fbioe.2021.779353, ISSN2296-4185.