Enhancing gesture recognition systems: integrating deep learning with advanced flexible sensing technologies

Mengmeng Qin*
College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518061, Guangzhou, China

ABSTRACT

Gesture-based interaction represents one of the most intuitive and immediate methods for human communication with their surroundings. The role of gesture recognition technology is particularly significant in fields like sign language interpretation and human-computer interface. Flexible sensors, characterized by their high sensitivity, excellent stretchability, and affordability, are particularly well-suited for incorporation into wearable devices. This paper provides a comprehensive overview of gesture recognition systems developed in recent years that utilize flexible sensors. It delves into the hardware architecture of notable data gloves and emerging wrist-worn devices. The study employs traditional machine learning techniques for recognizing both static and dynamic gestures. Furthermore, it explores the creation of advanced deep learning models, utilizing Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) architectures. Additionally, the paper examines the potential future applications of gesture recognition systems, highlighting their utility in smart home environments and surgical training. In conclusion, the article identifies existing challenges in terms of environmental robustness of sensors, signal latency, and the need for enhanced data set development.

Keywords: Gesture recognition, flexible pressure sensor, machine learning, deep learning

1. INTRODUCTION

Gesture recognition represents a pivotal component in AI systems, essential for interpreting human movements and extracting meaningful signals[1]. This technology, crucial in fields such as sign language translation, human-computer interaction, medical diagnosis, and smart transportation, primarily focuses on recognizing human body language. The human hand, a complex articulated object comprising 27 bones and 5 fingers, each with three joints, presents a significant challenge in gesture recognition. Before recognition can occur, hand modeling is categorized into temporal and spatial dimensions. Spatial modeling, often employing visual sensors, captures gestures intuitively, extracting features such as geometry, color, and texture from images for recognition. However, the complex structure of hand joints, their degrees of freedom, and variations in size, shape, and color, coupled with environmental influences on image quality, render both static and dynamic gesture modeling challenging.

On the other hand, temporal modeling typically utilizes data gloves equipped with flexible sensors. These gloves offer advantages such as low latency, high accuracy, minimal environmental interference, and consistent performance across different user demographics. They work by collecting sequential signals related to the bending of knuckles for recognition purposes. A fundamental approach to gesture recognition involves setting manual thresholds for each sensor to differentiate between bending and pointing states. However, the extensive number of finger joints and sensors significantly complicates signal processing. With the advent of artificial intelligence, sign language translation systems leveraging machine learning or deep learning algorithms have become capable of recognizing a broader range of sentence combinations in real-time, enhancing efficiency and accuracy in translation.

Gesture signals can be effectively classified using traditional machine learning algorithms like decision trees, support vector machines (SVMs), naive Bayes, or hidden Markov models (HMMs). However, these traditional algorithms often rely on heuristic manual feature extraction, leading to superficial feature utilization and limitations in model accuracy and generalizability. In contrast, deep learning excels in unsupervised learning, eliminating the reliance on costly labels

*2021280519@email.szu.edu.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133953J · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048708

and reducing the burden of feature extraction. It automatically learns more sophisticated and relevant features through neural networks, significantly benefiting real-time gesture recognition.

2. HARDWARE INNOVATIONS IN GESTURE RECOGNITION

The bending of finger joints and the deformation of the skin and tendons under the wrist skin contain a large amount of information about gesture changes, which can be obtained using flexible pressure sensors. Flexible pressure sensors have the characteristics of high sensitivity and fast response speed, and are now popularly used in real-time gesture recognition. In addition, it is mechanically robust, has good extensibility, is light in weight, fits the human skin, and is suitable for users to wear daily for a long time. Flexible pressure sensors can be divided into piezoresistive, piezoelectric and capacitive pressure sensors. Among them, the piezoresistive type converts changes in force into changes in resistivity. The conversion mechanism is relatively simple and low-cost for system design, so it is considered the most common one. In addition, flexible sensors usually improve sensor performance through reasonable doping and synthesis of carbon nanomaterials, polymer materials, and hydrogel materials[2]. Figure 1 is the system architecture of gesture recognition based on the flexible sensors.

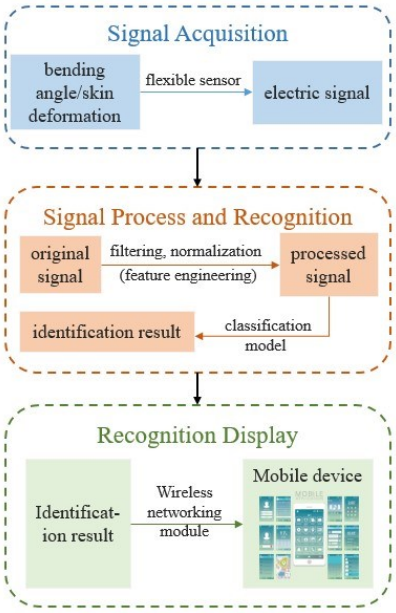

Figure 1. Flexible sensor-based gesture recognition system process.

3. DATA GLOVES

3.1 Sensing mechanism of flexible pressure sensor

(1) Capacitive Type

Resistive pressure sensors convert pressure changes into capacitance changes. The capacitive pressure sensor consists of an electrode and a dielectric layer, and the capacitance value C is defined as

$$C = \frac{\varepsilon_0 \varepsilon_r A}{d} \tag{1}$$

where, ε_0 is the vacuum permittivity, ε_r is the relative permittivity, A is the effective area of the electrode, and d is the distance between the plates. In the process of finger movement, movement changes are easily affected by pressure, so the pressure can be indirectly measured according to the change of capacitance signal.

(2) Resistive type

Resistive pressure sensors convert pressure changes into changes in resistance or current. The resistance is defined as

$$R = \frac{\rho L}{S} \tag{2}$$

where ρ is the resistivity, L is the length, and S is the cross-sectional area. The reasons for the change of resistance may include: the change of material shape, which leads to the change of L and S; Changes in the energy band structure of materials, some materials such as graphene stretch more than 20%, the energy band structure will change, affecting the resistivity of the resistance change; The contact resistance between materials changes, and some sensors exploit the closeness of the upper and lower layers of materials, resulting in resistance changes.

(3) Piezoelectric type

The piezoelectric pressure sensor converts the pressure signal into a voltage signal, and its induction mechanism comes from the piezoelectric effect of the piezoelectric material. The piezoelectric properties of piezoelectric materials are described by the piezoelectric constant, which reflects the ability of piezoelectric materials to convert mechanical energy into electrical energy. The larger the piezoelectric constant is, the better the piezoelectric performance is.

3.2 Data gloves

The data glove is composed of a central processing unit and sensors that capture gesture change signals. As shown in Figure 2, the data glove based on the flexible pressure sensor has sensors at the finger joints to measure the joint angles of the MCP and PIP joints of the fingers (DIP for the thumb). The wireless FPC board on the back of the hand performs filtering, amplification, digital-to-analog conversion and other operations on the sensor signal, and then transmits the signal to the computer through the Bluetooth module. Li et al.[3] embedded ten soft sensors in the data glove to measure the MCP and PIP joint angles of the five fingers. The data glove designed by Li et al. set 5 MWCNT flexible strain sensors at the joints. It is low-cost and has a total weight of 35.7 g. On average, Response time 2.173 ms[3].

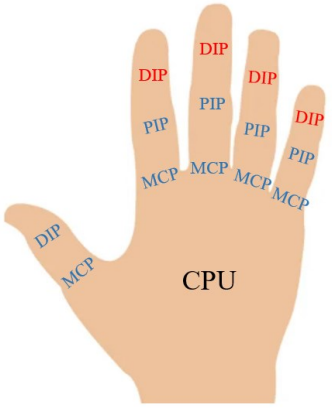

Figure 2. Schematic diagram of data glove.

3.3 Wrist-mounted devices

During the movement of gesture changes, the tendons and muscle groups of the hand movement will cause the skin of the wrist to deform. The wrist-worn gesture recognition device uses a flexible pressure sensor to obtain wrist deformation signals to avoid hindering the movement of the fingers. EMG is the most common and mature wearable gesture recognition method for wrist-worn devices while wrist-worn devices using flexible sensors have been developed and explored in recent years. Wang et al. used the ion electronic sensing method to design an ion electronic capacitor array with 4×8 sensing channels on the palm side of the flexible wristband, which has high sensitivity, good immunity, and small crosstalk between channels[4].

4. HAND GESTURE RECOGNITION ALGORITHMS

Gesture recognition can be viewed as a classification task, and a gesture recognition system with flexible sensors as hardware receives time series data. As shown in Figure 3, the sensors on each finger will generate waveforms of different strengths after being bent to different degrees, while unbent fingers will not. The flexible pressure sensor

converts the bending angle of fingers into the value of resistance, capacitance or voltage. Each gesture can be regarded as an n-dimensional vector input data.

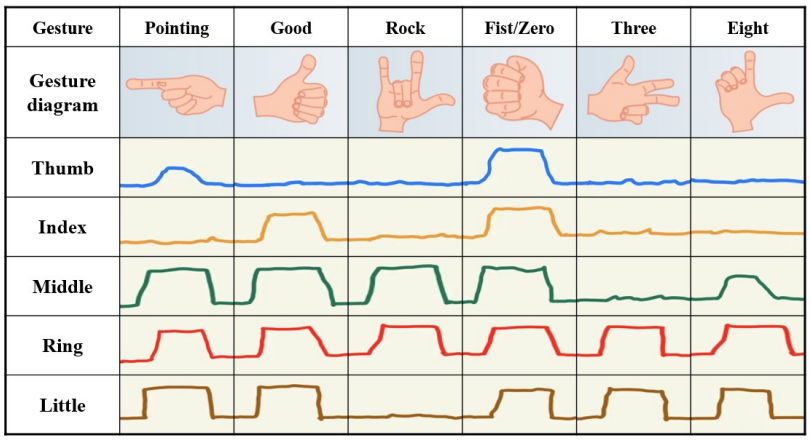

Figure 3. Waveform of flexible sensors on each finger in different gestures.

Before training the model, the data needs to be preprocessed and feature engineered. In static gesture recognition, filtering and normalization are usually used for data preprocessing. For example, a low-pass filter can be used to remove noise from the signal in the frequency domain. There are two types of normalization: min-max normalization, which maps features to a certain range to reduce sensor values from different sensing modes; z-score normalization, which shifts features to a similar distribution to shorten Training convergence time. Figure 4 specifically shows the process of signal processing and recognition in the system of Figure 1. After preprocessing, input data into the trained classifier. The classifier is a classification model composed of optimal parameters. In the task of gesture recognition, the classification accuracy and discrimination efficiency are usually used as the evaluation indicators of classification.

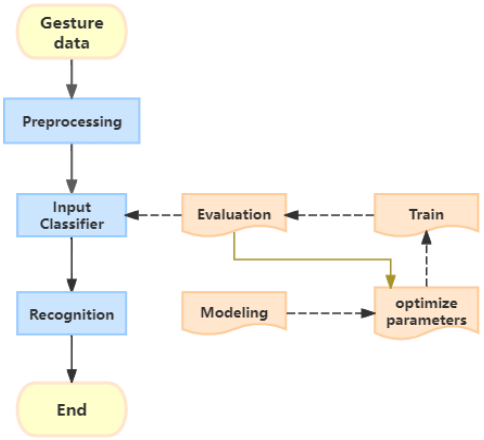

Figure 4. The process of signal processing and recognition.

4.1 Machine learning algorithms

In traditional machine learning methods, feature engineering is handled manually and can be divided into feature creation, feature extraction and feature selection. Feature creators use prior knowledge of hand activities to combine existing features or calculate each other to obtain new features. Principal component analysis (PCA) is often used for feature extraction in gesture recognition, which not only enhances the correlation with the target label, reduces the data dimension, but also eliminates redundancy and data noise. Feature selection methods include filters, such as variance filtering and correlation filtering, embedded and wrapper.

The preprocessed data is used as the input of the machine learning model and the classification results are output. In gesture recognition, commonly used machine learning models include KNN, SVM, LDA, hidden Markov model, etc.

Wu et al. used the K-NearestNeighbor (KNN) algorithm to achieve the highest test accuracy of 98.2% and the response time of less than 1s on 14 types of gestures with 750 samples[5].

4.2 Deep learning algorithms

Due to the advancement of big data and computing hardware, deep neural networks have been widely used in real-time gesture recognition in recent years. It has high accuracy and fast recognition speed for real-time gesture recognition.

Neural network is an end-to-end learning. It does not require manual feature extraction. It only needs to input the preprocessed data into the model, and it can be used to complete the task of feature extraction and classification of gesture timing at the same time. Neural networks with more than three layers are called deep neural networks. Compared with traditional machine learning, this nonlinear model can learn more advanced features from data.

Dynamic gesture recognition requires segmenting meaningful gestures from the continuous data stream and using the sliding window as the basic unit to be recognized by the model. For example, Lee, M and others proposed an algorithm based on gesture progression sequence (GPS), which divides the expression of gestures into three stages: preparation, nucleus and retraction, and sets a threshold, using a deep neural network to determine the start of a finger. and ends to segment the sliding window[6]. Since in real-time recognition, objects may have complex shapes and changes, deeper features are very important for the accuracy of recognition. Deep neural networks are more complex than simple machine learning methods. However, the network model can be simplified through quantification, pruning and other technologies, reducing the calculation amount and memory usage of the model, thereby improving the speed of gesture recognition.

(1) CNN

CNN, convolutional neural network, its basic structure includes Convolutional Layer: extract the features of the input data through convolution operation; Pooling Layer: reduce the spatial size of the feature map, reduce the amount of calculation, while retaining important information; Fully Connected Layer: pool the features output by the layer are mapped to the output categories. The convolutional layer of CNN can automatically learn features in images. By stacking multiple convolutional layers, the network can gradually abstract higher-level features, helping to improve the accuracy of gesture classification.

(2) LSTM

LSTM, standing for “Long Short-Term Memory”, is a variant of RNN. It can capture temporal correlations and long-term dependencies in gesture sequences, allowing LSTM to more accurately identify gesture actions or predict gestures. LSTM remembers through cells, and the cells decide to retain or forget the previous state. LSTM contains four interactive layers, three sigmoid layers and a tanh layer. Among them, the sigmoid layer compresses the value within the range of [0,1], so that the information multiplied by 0 will be forgotten, and the information multiplied by 1 will be saved. LSTM protects and controls the cell state through three gate structures: forget gate, input gate and output gate.

Lee and others’ model consists of two LSTM layers, six fully connected layers, and an output layer, their cross-validation results show that the accuracy of the training set is 100%, and the accuracy of the validation set is 98.5%[6]. Mutegeki et al. combined LSTM with CNN to achieve a recognition accuracy of 94.18%[7]. Yang and others used the LSTM unit combined with the attention mechanism for training. The bending prediction curve is consistent with the real data curve and reduces the communication delay in VR[8].

5. APPLICATIONS OF HAND GESTURE RECOGNITION

With the popularity of human-computer interaction, human gesture recognition technology is becoming more and more important. If the gesture recognition system based on flexible pressure sensor is combined with tactile sensor, it can provide a more realistic experience for users in augmented reality games and metaverse fields. In the field of unmanned driving, it can be applied to the gesture recognition of traffic police. In the field of Internet of Things, gestures can be used to control smart homes, robots, robotic arms, etc.

Human gesture recognition belongs to Human activity recognition (HAR). The collection, transmission, processing, analysis and display of finger joint flexion signals can also be applied to the recognition of other human joint activities, such as wrist, elbow, neck, etc. Therefore, it can not only be applied to help the deaf people to interpret sign language,

but also can be applied to the surgical training of surgeons, posture training of athletes, surgical medical diagnosis and other fields.

6. FUTURE PROSPECTS

In the actual use of flexible pressure sensor, it is often faced with a variety of different force environments, and it is necessary to ensure the stability of signal acquisition in the process of finger movement. The motion of the joint causes the electronic device attached to its surface to bear large tensile strain, which requires high flexibility of the material. To prepare stretchable sensors, materials with good stretchability are required. In addition, although some materials can be stretched, their performance is not stable after stretching. For example, the conductivity of electrodes composed of metal films and flexible substrates will drop sharply after being stretched to a certain extent.

Hysteresis and signal drift with time and environmental parameters remain two common non-ideal characteristics of resistive-structured flexible strain sensors. While materials engineering efforts can be made to improve the robustness of strain sensors to different environmental conditions, deep neural network structures are also changed to compensate on the system[9].

For gesture recognition whose hardware is a flexible sensor, the training data set is scarce. Gesture recognition uses ASL as the standard. For different sensor systems, it needs to collect corresponding time series data for training. The amount of data is small and the cost is high[10]. Data enhancement can be used to perform rotation, translation, scaling and other operations on existing data to expand the data set or transfer learning, and use model parameters pre-trained for other tasks to fine-tune or feature extraction.

7. CONCLUSION

This article presents a sophisticated gesture recognition system that utilizes flexible pressure sensors for hardware implementation, complemented by machine learning or deep learning techniques for accurate recognition. The system predominantly employs data gloves or wristbands as primary devices, incorporating sensors to capture nuances in finger flexion or skin deformation. These signals are first processed by a central processor and then fed into a computational model for precise gesture identification. In the realm of traditional machine learning, the approach is largely dependent on manual feature engineering, primarily suited for static gesture recognition. However, the advent of deep learning, particularly the utilization of Convolutional Neural Networks and Long Short-Term Memory networks, has revolutionized gesture recognition technology. These neural networks are adept at autonomously extracting complex, layered information, essential for dynamic gesture analysis. Enhancements in network accuracy, robustness, and generalizability are achieved through innovative combinations of CNNs and LSTMs, or by integrating attention mechanisms into these architectures. Such advancements in gesture recognition have broad implications across various domains, including human-computer interaction, medical diagnostics, and linguistic communication.

REFERENCES

[1] Faisal, M. A. A., Abir, F. F. and Ahmed, M. U., “Sensor dataglove for real-time static and dynamic hand gesture recognition,” In 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 1-7 (2021).

[2] Nan, X., Wang, X., Kang, T., Zhang, J., Dong, L., Dong, J., Wei, D., et al., “Review of flexible wearable sensor devices for biomedical application,” Micromachines 13(9), 1395 (2022).

[3] Li, Y., Yang, L., He, Z., Liu, Y., Wang, H., Zhang, W., Song, G., et al., “Low-cost data glove based on deep-learning-enhanced flexible multiwalled carbon nanotube sensors for real-time gesture recognition,” Advanced Intelligent Systems 4(11), 2200128 (2022).

[4] Wang, T., Zhao, Y. and Wang, Q., “A flexible iontronic capacitive sensing array for hand gesture recognition using deep convolutional neural networks,” Soft Robotics 10(3), 443-453 (2023).

[5] Wu, X., Luo, X., Song, Z., Bai, Y., Zhang, B. and Zhang, G., “Ultra-robust and sensitive flexible strain sensor for real-time and wearable sign language translation,” Advanced Functional Materials 33, 2303504 (2023).

[6] Lee, M. and Bae, J., “Deep learning based real-time recognition of dynamic finger gestures using a data glove,” IEEE Access 8, 219923-219933 (2020).

[7] Mutegeki, R. and Han, D. S., “A CNN-LSTM approach to human activity recognition,” In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 362-366 (2020).
[8] Yang, L. I., Jiang, J., Wang, R., Mao, Z., Fang, L., Qi, Y. and Yu, H., “Fully flexible smart gloves and deep learning motion intention prediction for ultra-low latency VR interactions,” IEEE Sensors Letters 7, 1-4 (2023).
[9] Si, Y., Chen, S., Li, M., Li, S., Pei, Y. and Guo, X., “Flexible strain sensors for wearable hand gesture recognition: from devices to systems,” Advanced Intelligent Systems 4(2), 2100046 (2022).
[10] Donati, M., Vitiello, N., De Rossi, S. M. M., Lenzi, T., Crea, S., Persichetti, A. and Carrozza, M. C., “A flexible sensor technology for the distributed measurement of interaction pressure,” Sensors 13(1), 1021-1045 (2013).