Journal of Computer and Communications, 2024, 12, 103-114
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Gesture Recognition Based on Time-of-Flight Sensor and Residual Neural Network

Yuqian Ma, Zitong Fang, Wen Jiang, Chang Su, Yuankun Zhang, Junyu Wu, Zhengjie Wang*

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, China

Email: 2206911173@qq.com, qin342442984@qq.com, 2334620367@qq.com, 3190213254@qq.com, 3072164579@qq.com, youngwu1027@163.com, *cieewangzj@163.com

How to cite this paper: Ma, Y.Q., Fang, Z.T., Jiang, W., Su, C., Zhang, Y.K., Wu, J.Y. and Wang, Z.J. (2024) Gesture Recognition Based on Time-of-Flight Sensor and Residual Neural Network. *Journal of Computer and Communications*, **12**, 103-114.
https://doi.org/10.4236/jcc.2024.126007

Received: May 7, 2024
Accepted: June 24, 2024
Published: June 27, 2024

Abstract

With the advancement of technology and the increase in user demands, gesture recognition played a pivotal role in the field of human-computer interaction. Among various sensing devices, Time-of-Flight (ToF) sensors were widely applied due to their low cost. This paper explored the implementation of a human hand posture recognition system using ToF sensors and residual neural networks. Firstly, this paper reviewed the typical applications of human hand recognition. Secondly, this paper designed a hand gesture recognition system using a ToF sensor VL53L5. Subsequently, data preprocessing was conducted, followed by training the constructed residual neural network. Then, the recognition results were analyzed, indicating that gesture recognition based on the residual neural network achieved an accuracy of 98.5% in a 5-class classification scenario. Finally, the paper discussed existing issues and future research directions.

Keywords

Hand Posture Recognition, Human-Computer Interaction, Deep Learning, Gesture Datasets, Real-Time Processing

1. Introduction

With technology continuously advancing and demands increasing, research on human activity recognition has gained widespread interest. Within the domain of human activity recognition, human gesture recognition has garnered significant attention, particularly in the field of human-computer interaction, owing to its broad application prospects [1]. Human gesture recognition can be categorized based on different signals.

1.1. Gesture Recognition Technologies

1) Gesture recognition based on vision and wireless signals

Currently, gesture recognition using visual sensors mostly involves extracting features from images or video streams [2]. For example, Arijit Das *et al*. [3] developed a visual-based gesture recognition system utilizing ASL as the language dataset and YOLO-v5 as the primary algorithm model. The overall accuracy reached 93% during training.

Meanwhile, some authors implement gesture recognition using FMCW radar signals. For instance, Wang *et al*. proposed a gesture recognition based on FMCW radar [4], which collects signals using radar and analyzes them through signal processing. In 2019, Wang *et al*. [5] proposed a TS-FNN method, improving average recognition accuracy by about 5% compared to traditional CNNs.

Visual sensor data requires complex processing and is sensitive to lighting, while wireless signal-based recognition often has lower accuracy due to environmental influences and complex signal fluctuations.

2) Gesture Recognition Using a Time-of-Flight Device

Time-of-flight-based gesture recognition primarily utilizes the product of laser emission time difference and the speed of light to determine the distance light travels through the air. This method uses the return time to obtain distance information about the object, thereby detecting the object's shape and recognizing gestures. This technique is applicable in many fields, including exhibition identification, security monitoring fields, medical domains, and more [6].

An *et al*. [7] propose a hand gesture recognition using a Time-of-Flight (ToF) device. Gesture recognition is accomplished by employing a Time-of-Flight (ToF) camera and deep neural networks that process 3D point cloud data. The networks achieve an average recognition accuracy exceeding 99%, with the PointNet++ network exhibiting superior performance compared to other architectures. Wang *et al*. [8] propose a dynamic hand recognition system using 8×8 ToF sensors. A compact parallel convolutional network, dubbed TPC-Net, is developed for dynamic gesture recognition. The proposed system demonstrates an impressive average accuracy of 98.47% across six dynamic gestures, as evidenced by experimental results. Moreover, the system outperforms existing methods that utilize low-pixel Time-of-Flight (ToF) sensors while maintaining a lower model complexity.

ToF gesture recognition systems exhibit high accuracy and, compared to vision-based recognition, are less affected by environmental variations, more effectively preventing privacy breaches, and reducing the challenges posed by signal fluctuations typical of wireless signal-based methods.

1.2. Hand Gesture Recognition Algorithm

Gesture recognition can generally be regarded as a classification system, allowing the implementation of gesture recognition through various classification methods. Among these algorithms, deep learning and machine learning applica-

tions are the most widespread. These technologies can utilize multiple types of data signals, including vision, wireless, and laser signals, presenting a broad spectrum of potential applications [9].

1) Machine learning algorithm

Machine learning-based gesture recognition technology trains models to identify human gestures. For example, Wong *et al*. [10] utilized Error-Correcting Output Codes Support Vector Machine (ECOC-SVM) and k-Nearest Neighbors (KNN) classifiers, achieving a 99% recognition rate with KNN and normalized feature data, and 97% with ECOC-SVM. Additionally, Moin *et al*. [11] developed a gesture recognition system using machine learning and wearable technology, achieving 97.12% accuracy for individual gestures tested with 13 different gestures. Even with 21 gestures, the system maintained a high accuracy of 92.87%. These results highlight how machine learning enhances gesture data analysis accuracy.

2) Deep learning algorithm

Deep learning is a type of machine learning that utilizes multi-layer neural network structures to learn and extract high-level features from data, thus enabling the identification and categorization of complex patterns. Deep learning is extensively applied in the field of gesture recognition, fault tolerance, high parallelism, and resistance to interference. Gong *et al*. [12] experimented with gesture recognition using mmWave radar, enhanced by deep learning techniques. They adapted the typical VGG16 network into a 3D-VGG16-NET to achieve an impressive accuracy of 99.38%. Additionally, Sun *et al*. [13] improved the accuracy of gesture image recognition through deep learning. They optimized parameter selection and implemented convolutional and fully connected network structures, along with algorithms for fingertip detection and elimination, to enhance recognition precision. The robust data-fitting capabilities of deep learning substantially improve gesture recognition accuracy, making it a valuable tool in domains like smart homes, virtual reality, and healthcare, showing a clear promise for future research.

This paper explores a gesture recognition system that utilizes the Time-of-Flight sensor VL53L5 and the ResNet50 residual neural network. Specifically, the system leverages the VL53L5 sensor and ResNet50 network, processing perceptual data through preliminary treatments before feeding it into the well-constructed residual network to facilitate gesture recognition.

The contribution of this paper can be described as follows.

1. We build the system using the VL53L5 sensor for gesture recognition. This sensor utilizes time-of-flight technology to acquire distance data in an 8x8 grid format. By analyzing this data, it accurately captures the spatial positions and motion trajectories of gestures, thereby providing reliable data support for subsequent gesture recognition tasks.

2. We use ResNet to implement the system, it proved the general performance of a simple model for hand recognition. This model demonstrates outstanding

performance in handling gesture data, effectively extracting and learning abstract features of gestures. Consequently, it achieves precise recognition and classification of gestures. This research validates the applicability of this neural network in behavior recognition based on ToF sensors, showcasing significant potential for broader application.

The structure of this paper can be outlined as follows. Section 1 provides a background review, introducing the current state of research in gesture recognition based on vision, millimeter-wave radar, and laser technologies, and summarizing the main research content of this paper. Section 2 introduces the framework of the gesture recognition system, including the gesture recognition scheme and principles, experimental equipment and its working principles, and the neural network structure. Section 3 details the experimental design of this paper, covering data collection and processing, and analyzing the experimental results and system accuracy. Section 4 discusses the current system's limitations and outlines future research directions. Section 5 concludes the paper by summarizing the applications and accuracy of the gesture recognition system based on the ToF sensor.

2. The Framework of the Gesture Recognition System

This system uses VL53L5 and residual network to implement the gesture recognition system, and the components and implementation principles of the system are introduced below.

2.1. Gesture Recognition Framework Based on ToF Sensor

The basic process of recognition of this gesture includes the following four steps, as shown in Figure 1.

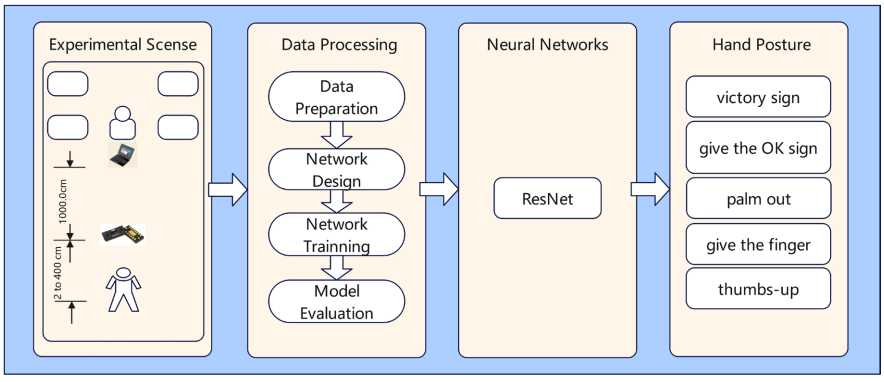

Figure 1. Gesture recognition scenarios and principles.

1) Data reception: We invite four students to collect actions respectively, each person collects five actions, including five static gestures such as thumbs up, "OK" and uses VL53L5 sensors to detect the posture of human hands, and transmit the detected data to the P-NUCLEO-53L0A1 development board.

2) Data processing: We train the ResNet neural network to ensure that its ac-

curacy meets our requirements, and the data collected by the sensor will be preprocessed and the gesture model will be constructed.

3) Data analysis: The trained ResNet neural network is used to extract temporal and spatial features by convoluting the video containing gestures, and the data from the VL53L5 sensor is systematically analyzed.

4) Display result: The neural network model of gesture recognition analyzes the incoming data to determine the meaning of the gesture, to give the recognition result.

2.2. The Experiment Device Description

The VL53L5 uses an avalanche photodiode pixel array to meet the single-photon detection, which directly measures the distance to the target according to the reflection time of the emitted photon, which can achieve accurate ranging regardless of the surface characteristics of the target. This solution offers accurate, high-speed ranging, low power consumption, competitive system cost, and easy integration through flexible mechanical design. Due to the small size of the VL53L5CX-SATEL breakout board, it can be more easily integrated into development and evaluation equipment, which has clear advantages as a next-generation imaging technology. It is less affected by environmental changes and other factors, and the collected data is more stable. Since the process only obtains the outline matrix of the target, the need for privacy security is met.

The VL53L5 mainly uses an 8 × 8 grid mode for ranging, which can provide up to 64 ranging areas for multi-object detection applications, with a maximum measurement distance of 4 meters in dark environments, and a wide rectangular field of view with a diagonal field of view of 65˚, realizing multi-distance and multi-area detection. Based on the use of the VL53L5 sensor, the reference SPAD array is also used to compensate for any time drift in the VCSEL optical signal. By measuring the time it takes for the emitted VCSEL pulse to return to the sensor, the VL535 sensor can determine the target distance, obtain 8*8 three-dimensional distances of the target, and obtain the target contour matrix, to collect the data correctly. The basic operating parameters are shown in **Table 1**.

Table 1. The parameter of VL53L5.

VL53L5 Working Parameter	Value
Sample rate	Up to 60 Hz
Infrared emitter	940 nm
Ranging	2 to 400 cm per zone
Operating voltage	IOVDD: 1.8 or 2.8 V or 3.3 V AVDD: 2.8 V or 3.3 V
I2C interface	I2C: 400 kHz to 1 MHz serial bus, address: 0 × 52
Operating ranging mode	Continuous or Autonomous

2.3. The Structure of Residual Neural Networks

A key feature of ResNet is its use of residual joining, a connection mechanism that allows the network to learn the residual mapping between the inputs and outputs of each layer, rather than directly learning the complex mapping relationships, as shown in **Figure 2**. This feature enables the network to learn deep and complex features more efficiently, while effectively alleviating the gradient vanishing problem common in deep neural networks. Compared with traditional deep networks, ResNet50 performs better in terms of accuracy and stability [14]. Therefore, we chose a method based on the ResNet50 neural network to analyze the collected gesture results.

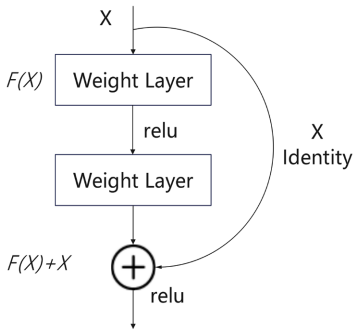

Figure 2. ResNet neural network.

We describe the basic residual structure. The input vector X is weighted and nonlinearized and summed with the original input, which retains the original information and avoids gradient explosion. In residual neural networks, the basic residual units are composed of convolutional layers, BN layers, ReLU layers, and jump connections. The convolutional layer performs convolution operations and extracts features from the input data. The BN layer normalizes the input data for each layer. The purpose of normalization is to increase the robustness of the network, reduce the sensitivity of the network to the input gesture data, accelerate the convergence speed of the network, and improve the generalization performance of the network. The role of the ReLU layer is to increase the flexibility of the model, accelerate the convergence process, avoid gradient vanishing, etc. Hopping connections can mitigate vanishing gradients. The structure of the residual element determines the performance of the network to a large extent [15].

In this paper, the ResNet50 neural network composed of multiple residual structures is used for gesture classification, and its structure is shown in **Figure 3**. During data processing, the first convolutional layer of the ResNet50 model accepts 1 input channel and 64 output channels. The convolution kernel size is (7, 7), the stride length is (2,2), and the fill is (3, 3). The input dimension of the fully connected layer of the ResNet50 model is 2048 and the output dimension is 5, a total of 19 layers.

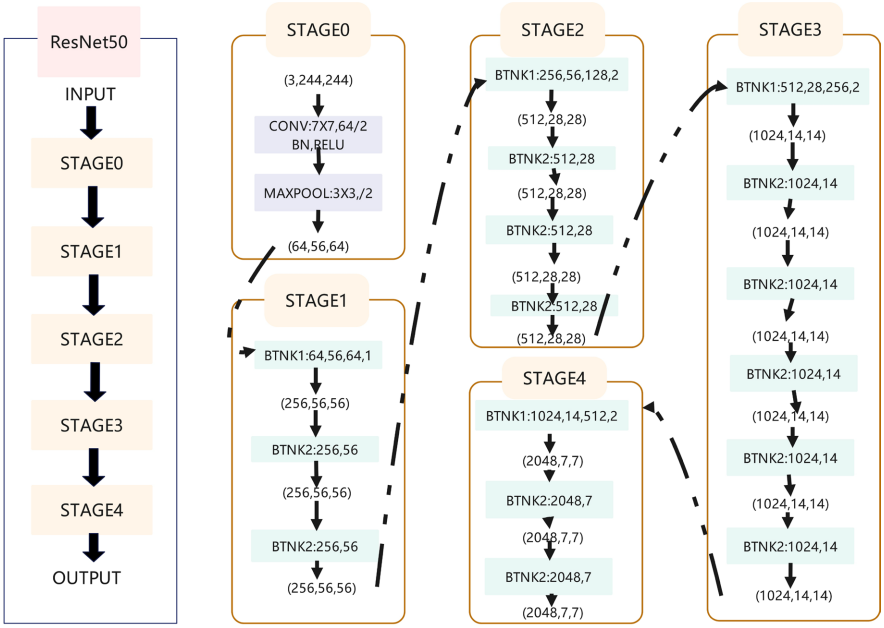

Figure 3. Structure diagram of the ResNet50 model.

The data processing process of the ResNet50 neural network for gesture recognition includes the following four steps:

1) Data preprocessing: This process aims to implement data cleaning and format conversion for input data. First, the collected gesture data is cleaned to remove possible abnormal or invalid data points to ensure the accuracy and reliability of the gesture data. The lattice data is then converted into a format suitable for neural network input. Data is rearranged, reconstructed, or encoded to ensure that it can be properly understood and processed by neural networks.

2) Data division: The preprocessed gesture dataset is divided into a training set, a validation set, and a test set, usually using 70% of the data as the training set, 20% of the data as the validation set, and the remaining data as the test set.

3) Build the model: Select the appropriate ResNet50 model structure, define the loss function, optimizer, and evaluation index of the model, and ensure that the model can effectively optimize the parameters and accurately evaluate the performance during the training process. The constructed gesture model is compiled, and each component is integrated to prepare for subsequent gesture recognition training and evaluation, to achieve higher accuracy and robustness.

4) Model training: The model is trained using the training set, and the model is validated and tuned through the validation set. In the training process, techniques such as learning rate decay and early stop strategy can be used to improve the training effect of the model. The test set is used to evaluate the trained model and the performance accuracy.

3. Experiment Setting and Result Analysis

3.1. Experiment Setting

1) Data collection parameter settings

This system uses a residual neural network to realize the gesture recognition system. The device is placed on the table, 0.5 meters away from the person, and performs 5 static actions, which are thumb up, Y-shaped, "OK" gesture, upright middle finger, and open palm. There were 4 participants, each of whom performed the above 5 movements, each of which was performed 30 times and 600 samples.

2) Neural network training parameter settings

The parameters of the network training are as follows: with an epoch of 30, a batch size of 16, and a learning rate of 0.0001, we processed five static gestures, the loss function uses cross-entropy, and the optimization function adopts the Adam method.

3.2. Experiment Result Analysis

According to the experimental results, we can see that the model's recognition performance of static gestures is relatively stable under these conditions. By training 30 epochs, the model has reached a certain convergence state, the batch size of 16 ensures that the sample data is fully utilized during each training process, and the learning rate of 0.0001 makes the model more stable in the training process and avoids the problem of gradient explosion or gradient vanishing.

The training process and training results are shown in Figure 4 and Figure 5. The loss function is used to evaluate the difference between the built model and the actual value, and the update function is used to update the records in the database. Figure 4 shows that the model fitting effect is good, and the difference between the predicted value of the model and the actual value gradually decreases and tends to 0. The classification results of the model for the five static gestures are shown in Figure 5. It shows that the model has a good recognition effect on different gestures, with an accuracy of 98.5. This shows that the neural network structure and training parameters we selected can effectively classify and recognize static gestures, which provides a good foundation for subsequent applications.

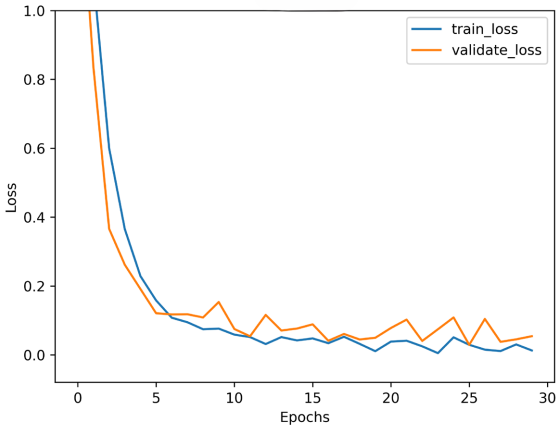

Figure 4. Change of gesture data processing loss function.

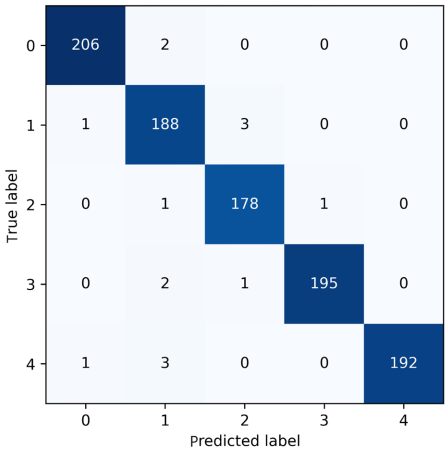

Figure 5. Gesture data processing confusion matrix.

We compare the result of this paper with other gesture recognition to evaluate the system's performance. Since few papers implement gesture recognition using VL53L5, we have to compare this paper with other visual sensors and wireless sensors. Compared to visual sensors, this system performs better in recognizing gestures in low-light environments and can be applied in a wider range of experimental settings [16]. Compared to millimeter-wave radar, this system has a high refresh rate and quick response, ensuring strong real-time performance [17].

4. Discussion

We discuss the limitations of this paper here, mainly including the lack of research on multi-person gestures and complex dynamic gesture recognition. Future research directions mainly consider the method based on this paper, which can realize complex three-dimensional human posture and emotion recognition.

4.1. System Limitations

1) Dataset limitations

Although we collected a large amount of gesture data for training and testing, there are still limitations in the diversity and coverage of the dataset. The quality and quantity of the dataset directly impact the performance and generalization ability of the system. If the dataset lacks samples of certain specific types or scenarios of gestures, the system may fail to effectively recognize these gestures.

2) Model generalization

Although we selected the ResNet neural network as the gesture recognition model and trained and optimized it, the model still exhibits certain limitations in terms of generalization. In practical applications, various factors such as changes in lighting conditions, differences in the speed and magnitude of gesture movements, etc., which may differ from the training data, could impact the model's recognition performance.

3) Multi-person scenarios and dynamic gesture

In this study, we primarily analyzed single-person gesture recognition results. However, in multi-person scenarios, the complexity and volume of data increase significantly, complicating processing and analysis. The system can easily misclassify overlapping gestures, requiring more flexible and intelligent algorithms. At the same time, dynamic gesture recognition involves continuous changes in hand postures and movement trajectories. Accurate processing needs to consider temporal sequences and spatial relationships. Thus, for continuous actions, a neural network with time series capabilities, like RNN or LSTM, is necessary to track changes in hand postures and motion trajectories over time.

4.2. Future Research Directions

1) Three-dimensional posture recognition of the human body

In this paper, a gesture recognition system is implemented by using a neural network, which can be extended to human daily activity recognition and human posture recognition. Although the data obtained by this sensor is relatively scarce, it is possible to accurately estimate the posture of the human body by extracting the features of the neural network and using prior knowledge of human activities, such as the fact that the length of the limbs does not change, and the direction and change of movement are affected by the joints.

2) Application of emotion recognition

Human emotion recognition is an important research content in human-computer interaction, and the processing of human expression data obtained by ToF sensors through neural networks can realize human expression recognition, which provides a research basis for human-computer interaction and emotion recognition applications. By constructing a neural network, extracting facial features, and analyzing feature changes, this method can quickly and accurately capture signal features and achieve fine facial expression changes, which expands the application scope of this study.

5. Conclusion

Gesture recognition is an important research content in human-computer interaction, and gesture recognition based on time-of-flight sensors has gradually become a hot research topic. In this paper, the VL53L5 ToF sensor is taken as the research object, and a gesture recognition system based on VL53L5 sensor and ResNet50 neural network is realized by collecting gesture data. Through this system, we can accurately recognize and classify user gestures to the control and operation of a device or system. In the process of experimentation, we collected a large amount of gesture data, preprocessed and extracted features, and selected the ResNet50 neural network as the gesture recognition model, which was trained and optimized. Through a large number of experimental verifications and testing, we found that the constructed ResNet50 neural network performed well in recognizing gestures, with 98.5 accuracy. Through this technology, we

can realize gesture control of devices or systems, creating a more convenient and intelligent interactive experience.

Fund

The work is funded by the foundation of the Innovation and Entrepreneurship Training Program for College Students (202310424138) and Exploration and Practice of Programming Course Teaching Models for Electronic Information Engineering Major in the Context of New Engineering Education (QX2022M39).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Guo, L., Lu, Z.X. and Yao, L.G. (2021) Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review. *IEEE Transactions on Human-Machine Systems*, **51**, 300-309. https://doi.org/10.1109/THMS.2021.3086003

[2] Tripathi, R. and Verma, B. (2023) Survey on Vision-Based Dynamic Hand Gesture Recognition. *The Visual Computer*. https://doi.org/10.1007/s00371-023-03160-x

[3] Das, A., Maitra, K., Roy, S., Ganguly, B., Sengupta, M. and Biswas, S. (2023) Development of a Real Time Vision-Based Hand Gesture Recognition System for Human-Computer Interaction. *Proceedings of the* 2023 *IEEE* 3*rd Applied Signal Processing Conference* (*ASPCON*), Haldia, 24-25 November 2023, 294-299. https://doi.org/10.1109/ASPCON59071.2023.10396583

[4] Wang, Z.J., Liu, F., Li, X., Ma, M.J., Feng, X.X. and Guo, Y.J. (2023) A Survey of Hand Gesture Recognition Based on FMCW Radar. *Proceedings of the* 8*th International Conference on Communication and Information Processing*, Beijing, 3-5 November 2022, 73-79.

[5] Wang, Y., Wang, S.-S., Tian, Z.-S., Zhou, M. and Wu, J.-J. (2019) Two-Stream Fusion Neural Network Approach for Hand Gesture Recognition Based on FMCW Radar. *Acta Electronica Sinica*, **47**, 1408-1415.

[6] Kai, Z.Q. and Shi, J.F. (2022) Modeling of Laser Femtosecond Gesture Recognition Algorithm and Implementation of Simple Navigation Microsystem. *Computer Measurement and Control*, **30**, 246-254.

[7] An, D., Xu, T.X., Zhang, Y.W. and Yue, Y. (2022) Hand Gesture Recognition Using ToF Camera and 3D Point Cloud Networks. *Proceedings Frontiers in Optics + Laser Science* 2022 (*FIO*, *LS*), Rochester, 17-20 October 2022, JW4B.56.

[8] Wang, X.T., Feng, W.W., Shi, Z.G. and Wang, Y. (2023) Research on Dynamic Gesture Recognition with Low-Pixel ToF-Sensors. *Proceedings of the* 2023 *International Conference on Ubiquitous Communication* (*Ucom*), Xi'an, 7-9 July 2023, 150-155. https://doi.org/10.1109/Ucom59132.2023.10257656

[9] Liu, W.B. and Liu, J. (2020) Overview of Vision-Based Dynamic Gesture Recognition. *Computer Applications and Software*, **37**, 190-197.

[10] Wong, W.K., Juwono, F.H. and Khoo, B.T.T. (2021) Multi-Features Capacitive Hand Gesture Recognition Sensor: A Machine Learning Approach. *IEEE Sensors Journal*, **21**, 8441-8450. https://doi.org/10.1109/JSEN.2021.3049273

[11] Moin, A., Zhou, A., Rahimi, A., Menon, A., Benatti, S., Alexandrov, G., *et al*. (2021) A Wearable Biosensing System with In-Sensor Adaptive Machine Learning for Hand Gesture Recognition. *Nature Electronics*, **4**, 54-63. https://doi.org/10.1038/s41928-020-00510-8

[12] Gong, S.F., Fang, Y.M., Shi, H.Y., Yan, X.Y. and Wu, Z.F. (2023) Experimental Teaching Design of Millimeter Wave Radar Gesture Recognition Based on Deep Learning. *Experimental Technology and Management*, **40**, 168-176.

[13] Sun, B.W. and Yu, F. (2021) Dynamic Gesture Recognition and Interaction of Monocular Camera Based on Deep Learning. *Journal of Harbin University of Science and Technology*, **26**, 30-38.

[14] Gong, A.J. (2024) Research on Building a Classification Diagnosis Model of Eye Diseases Based on ResNet Deep Neural Network. *Journal of Medical Forum*, **45**, 379-383.

[15] Peng, H., Deng, X.Z., Niu, Y.X. and Liu, Y. (2024) Application of ResNet50 Model in Recognition and Classification of Pneumonia. *Journal of Fujian Computer*, **40**, 9-13.

[16] Chen, W.J. and Zhang, Z. (2019) Hand Gesture Recognition Using sEMG Signals Based on Support Vector Machine. *Proceedings of the* 2019 *IEEE* 8*th Joint International Information Technology and Artificial Intelligence Conference* (*ITAIC*), Chongqing, 24-26 May 2019, 230-234. https://doi.org/10.1109/ITAIC.2019.8785542

[17] Xia, Z.Y., Zhou, C.L., Jie, J.Y., Zhou, T., Wang, X.F. and Xu, F. (2020) Micro-Motion Gesture Recognition Based on Multi-Channel Frequency Modulated Continuous Wave Millimeter Wave Radar. *Journal of Electronics & Information Technology*, **42**, 164-172.