Embedded Multimodal Perception Smart Glasses: Innovation and Practice of Barrier-Free Interaction Technology

Xiangxuan Ji, Zhiyuan Li*
School of Artificial Intelligence and Software, Ke Wen College of Jiangsu Normal University, Xuzhou, Jiangsu, China
**Corresponding Author*

Abstract: Addressing the communication gap between Deaf and Hard of Hearing (DHH) individuals and the hearing population, as well as the numerous inconveniences faced by Blind and Visually Impaired (BLV) individuals in environmental perception and daily travel, this paper develops an intelligent glasses-based barrier-free interaction system leveraging embedded multimodal perception technology. This system deeply integrates cutting-edge technologies such as computer vision, intelligent audio analysis, tactile feedback control, and edge computing. It innovatively employs a lightweight spatiotemporal graph convolutional network architecture, multimodal information fusion algorithms, and optimized environmental perception models, successfully achieving core functionalities such as real-time dynamic sign language recognition, precise speech-to-tactile semantic conversion, rapid environmental hazard warning, and high-precision navigation. The system utilizes binocular vision and multi-sensor collaborative fusion solutions, effectively breaking through the performance bottleneck of traditional single-modality assistive devices and achieving low-power consumption (1.2W) and low latency (end-to-end delay < 500ms) multimodal information collaborative processing on an embedded hardware platform. Test results show that the dynamic sign language recognition accuracy reaches 87.6%, spatial positioning accuracy is up to 0.3 meters, and hazard warning response time is only 95ms. This system builds a comprehensive, barrier-free, and inclusive interaction environment for special groups, facilitating their deep integration into the digital society.

Keywords: Barrier-Free Interaction Technology; Multimodal Information Fusion; Dynamic Sign Language Recognition; Tactile Feedback Mechanism; Edge Intelligent Computing; VSLAM Navigation System

1. Introduction

1.1 Research Background

In the rapid development of the digital society, the deaf and dumb and visually impaired groups still face severe information interaction challenges: the deaf and dumb primarily rely on sign language for communication, but hearing individuals generally lack the ability to understand sign language, leading to blocked communication channels; due to the lack of visual information, the visually impaired face numerous safety hazards in daily scenarios such as environmental orientation and danger avoidance. Traditional assistive devices have obvious shortcomings: single-modal interaction cannot meet the needs of complex scenarios, they have poor adaptability in complex environments, and it is difficult to achieve a balance between lightweight design and real-time response. Most existing sign language recognition systems rely on heavy computing devices, making it difficult to meet the requirements of embedded deployment; visually impaired assistive devices mostly focus on a single navigation function and lack a multi-dimensional and comprehensive danger warning mechanism. Therefore, developing an intelligent assistive system with multi-modal interaction capabilities, lightweight characteristics, and high environmental adaptability has become the key to solving the information interaction barriers faced by special groups.

1.2 Research Objectives and Significance

This study aims to establish an efficient and seamless two-way barrier-free communication system among deaf and mute individuals,

visually impaired individuals, and able-bodied individuals. Through the innovative application of embedded multimodal perception technology, three core objectives are achieved:

Achieve real-time synchronous conversion between sign language and speech/text, completely breaking the communication barrier between deaf and mute individuals and hearing individuals;

Establish a voice-tactile semantic mapping mechanism to provide accurate and intuitive information feedback for the visually impaired;

Establish a multi-dimensional environmental perception and early warning system to comprehensively enhance the safety and travel convenience of visually impaired individuals.

The findings of this study not only facilitate special groups in participating in social interactions, accessing digital information, and deeply integrating into the digital society more conveniently, but also create significant inclusive social value and provide a new paradigm for the innovative development of barrier-free interaction technology.

1.3 Research Content and Structure

This paper focuses on two core areas: the integration optimization of multimodal interaction systems and the engineering implementation of lightweight technologies. Specifically, it covers the innovative design and implementation of dynamic sign language recognition modules, multimodal haptic interaction engines, and multimodal environmental perception modules for the visually impaired, as well as lightweight optimization schemes for hardware architecture and algorithm models. The subsequent chapters of the paper are arranged as follows: Chapter 2 provides an overview of the relevant technical background and research status; Chapter 3 elaborates on the overall system architecture design and implementation details of each functional module; Chapter 4 deeply analyzes the core technical principles and model training processes; Chapter 5 presents system performance test results and data analysis; and Chapter 6 summarizes the research findings and outlines future development directions.

2. Overview of Related Technologies

2.1 Basic theory of Convolutional Neural Network (CNN)

As a core deep learning model for processing grid-structured data, Convolutional Neural Networks (CNNs) leverage three key mechanisms: local connectivity, weight sharing, and pooling operations, enabling efficient extraction of hierarchical features from low-level to high-level data. A typical CNN architecture consists of an input layer, convolutional layers, activation layers, pooling layers, and fully connected layers: the convolutional layers extract features from the input data through sliding convolution kernels; the activation layers introduce nonlinear transformations (with the ReLU function being the current mainstream choice due to its effectiveness in addressing the vanishing gradient problem); the pooling layers achieve feature dimensionality reduction and translational invariance, enhancing the model's generalization ability; and the fully connected layers are responsible for feature integration and final output. During model training, the comprehensive application of Adam optimizer, Batch Normalization (BN), and Dropout regularization techniques can significantly improve the training efficiency and stability of CNNs, making them widely used in computer vision tasks such as image recognition and gesture detection [1].

2.2 Research progress in Multimodal Fusion Technology

Multimodal fusion technology can effectively enhance the system's adaptability to complex scenarios by integrating different types of modal information such as vision, hearing, and touch. This paper constructs two major technical systems: cross-modal perception conversion and intelligent environmental perception. The cross-modal perception conversion system is centered around a vibration tactile encoding matrix and utilizes an LSTM-CRF model to achieve context-aware conversion from speech to text to touch, ensuring accurate transmission of semantic information. The intelligent environmental perception system innovatively integrates an improved VSLAM algorithm and a dual-channel voiceprint recognition architecture to achieve collaborative operation of high-precision spatial positioning and rapid danger warning.

2.3 Analysis of the Current Status of Sign Language Recognition and Visually Impaired

Assistance Technology

Sign language recognition technology has evolved from early static recognition to dynamic recognition, and from monocular vision to multi-ocular vision. The Spatio-Temporal Graph Convolutional Network (ST-GCN) can effectively extract the spatio-temporal features of dynamic gestures by constructing a joint point graph structure, significantly improving recognition performance. In the field of visual impairment assistance technology, SLAM navigation and voiceprint recognition are two core research directions. The YAMNet model exhibits excellent performance in environmental sound classification tasks, accurately recognizing dangerous signals such as car horns and sirens. However, existing technologies still face many urgent issues to be addressed: the computational complexity of dynamic sign language recognition models is high, making it difficult to deploy them on embedded devices; the scarcity of localized sign language datasets leads to insufficient adaptability of models to local sign language variations; the balance between device power consumption and response delay has not been effectively resolved, all of which provide clear directions for improvement in this paper's research.

3. Overall System Design and Implementation

3.1 Hardware Platform Architecture Design

The smart glasses hardware platform adopts a lightweight wearable design concept, with core components including:

A binocular wide-angle camera, deployed on the side of the frame, supporting flexible angle adjustment, which can effectively expand the visual capture range and provide high-quality data input for sign language dynamic feature acquisition and environmental visual information acquisition;

A multimodal sensor module, integrating inertial sensors and high-precision audio sensors, to achieve deep fusion of visual-inertial data and precise acquisition of environmental sound;

A low-power edge computing unit, with powerful real-time data processing capabilities, optimizing the device's operating power consumption to 1.2W;

A voice interaction module, with a built-in miniature high-fidelity speaker, enabling clear voice output;

A tactile feedback module, integrating four linear motors at the temples, supporting 16-level intensity/frequency adjustment, capable of achieving diverse vibration feedback modes.

3.2 Software System Architecture Design

The software system adopts a modular design philosophy and consists of five core functional modules. The specific functions of each module are as follows:

3.2.1 Dynamic sign language recognition module

Based on the independently developed Lightweight Spatio-Temporal Graph Convolutional Network (LT-STGCN), this approach leverages the lightweight design strategy of improved YOLOv7-tiny and the depthwise separable convolution concept of MobileNetV2, combined with the principle of binocular vision, to achieve precise capture of hand dynamic features [2,3]. By utilizing the Leap Motion SDK to obtain real-time gesture joint point data, and employing a spatio-temporal feature parsing mechanism and adaptive skeletal modeling technology, it enables real-time and efficient parsing of complex continuous sign language. The model is trained on the WLASL dataset and incorporates a CoT Block to enhance temporal modeling capabilities, effectively improving the distinction between similar sign language actions.

3.2.2 Intelligent voice interaction module

The Mozilla DeepSpeech speech recognition model undergoes deep optimization, utilizing model compression technology to significantly reduce model storage footprint. An innovative approach is adopted by integrating directional noise reduction technology, combining beamforming algorithms with RNNoise deep learning denoising methods. This dual-pronged effort, focusing on both hardware signal processing and software algorithm optimization, notably enhances the accuracy and robustness of speech recognition. This module supports efficient speech-to-text conversion, laying a solid data foundation for sign language simultaneous interpreting and tactile encoding.

3.2.3 Multimodal Haptic Interaction Engine

Leveraging the unique advantages of RNN in temporal data processing, the LSTM-CRF model is employed to establish a precise mapping relationship between speech, text, and haptics. The system designs differentiated

vibration feedback modes based on different semantic types: short pulse vibration for ordinary conversations, continuous high-frequency vibration for emergency alarms, and double-pulse interval vibration mode for interrogative tone, achieving precise matching between semantic information and haptic feedback.

3.2.4 Multimodal Environment Perception Module for the Blind

The navigation function is achieved through a low-cost and cost-effective SLAM solution. By integrating visual and inertial data using an improved VSLAM algorithm, real-time spatial positioning with an accuracy of 0.3 meters is achieved. The environmental hazard perception function is based on a pre-trained YAMNet model, which can quickly identify dangerous signals such as vehicle horns and sirens, and trigger a vibration alert within 95ms. The system supports an offline deployment mode, ensuring rapid response even in a network-free environment [4].

3.2.5 Cross-modal perception and transformation system

With the vibration haptic coding matrix as the core foundation, the system deeply processes speech sequence information through the LSTM-CRF semantic understanding model, achieving contextual association conversion from speech to haptic coding [5]. The system supports the combined expression of multiple basic semantic units, enabling deep transformation from speech semantics to haptic coding, providing rich and accurate information feedback for visually impaired individuals [6,7]. The system architecture diagram is shown in Figure 1.

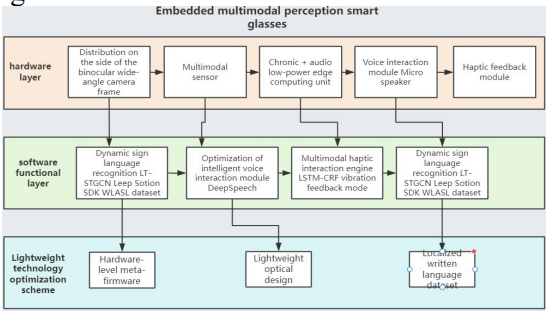

Figure 1. System Architecture Diagram

3.3 Optimization Scheme for Lightweight Technology

3.3.1 Hardware level optimization

By optimizing the sensor layout design and selecting low-power components, the overall power consumption of the device is strictly controlled at 1.2W. A lightweight optical design is implemented for the binocular wide-angle camera, which maximizes the reduction of device weight while ensuring imaging quality, thereby ensuring comfortable wearability [8,9].

3.3.2 Optimization at the algorithm level

By combining the DSConv efficient convolution operator with deep separable convolution technology, the speech recognition model and sign language detection model are compressed and optimized, reducing the model parameter count by more than 40%. Utilizing an edge computing architecture, end-to-end latency on mobile devices is less than 500ms, fully meeting real-time interaction requirements. At the same time, a localized dataset containing dialectal sign language is constructed, significantly enhancing the model's ability to recognize domestic sign language variations.

4. Technical Principles and Model Training

4.1 Cross-Modal Perception Transformation Principle

The core of the cross-modal perception transformation system lies in the construction of a speech-tactile semantic mapping mechanism. Initially, multi-dimensional feature extraction is performed on the speech signal. The temporal information of the speech sequence is processed through an LSTM network, and the coherence of semantic understanding is optimized using a CRF model. Subsequently, the speech semantics are precisely mapped to multi-dimensional tactile patterns. The vibratory tactile encoding matrix defines the correspondence between different semantic units and vibration parameters (intensity, frequency, duration), supports the combined expression of complex semantics, and achieves deep transformation from speech to tactile perception.

4.2 Principle of Intelligent Environmental Perception

The intelligent environmental perception system innovatively integrates an improved VSLAM algorithm with a dual-channel voiceprint recognition architecture: The improved VSLAM utilizes multi-sensor fusion technology to integrate visual, inertial, and other multi-source data, achieving high-precision real-time spatial positioning. The dual-channel voiceprint recognition architecture constructs a

comprehensive feature library of dangerous sound sources, employs a dual-channel parallel processing mechanism, and is capable of quickly identifying dangerous sound sources in the environment and issuing timely warnings. This system supports offline deployment and achieves low-latency multimodal information processing at the edge, ensuring rapid response capabilities in complex environments [10].

4.3 Visual Direction Implementation and Model Training

4.3.1 Data acquisition and preprocessing

The Leap Motion device is used to collect gesture data, and the coordinates and motion trajectories of hand joints are obtained in real-time through the API functions of the SDK. The voice data covers Chinese voice samples with different accents and speaking speeds, which are used for model training after preprocessing operations such as noise reduction and normalization. The dangerous sound data is obtained by combining public datasets with on-site collection, including various typical dangerous signals such as vehicle horns, alarms, and collision sounds.

4.3.2 Model training process

The dynamic sign language recognition model is trained based on the WLASL dataset, utilizing a spatiotemporal graph convolutional network (ST-GCN) to construct a graph structure. It maps the joint point relationships of gesture actions to the nodes and edges of the graph, extracting spatiotemporal features through convolution operations. During training, the Adam optimizer is employed, combined with batch normalization and Dropout regularization techniques, effectively preventing model overfitting. The speech recognition model is fine-tuned and optimized based on Mozilla DeepSpeech, enhancing real-time performance through model compression techniques. For dangerous sound recognition, the pre-trained YAMNet model is utilized, undergoing transfer learning on a custom dataset, significantly improving the accuracy of dangerous signal recognition.

5. Experimental Results and Performance Analysis

5.1 Experimental Environment and Evaluation Metrics

The experimental hardware environment utilizes an embedded platform for smart glasses (CPU: quad-core ARM Cortex-A53, GPU: Mali-T860, memory: 2GB), with Python 3.8 as the software environment and PyTorch 1.10 as the deep learning framework. The main evaluation metrics include: dynamic sign language recognition accuracy, end-to-end interaction delay, spatial positioning accuracy, hazard warning response time, and device operating power consumption.

5.2 Experimental Results

Sign Language Recognition Performance: On the WLASL dataset and the localized dialect sign language dataset, the dynamic sign language recognition accuracy reached 87.6%, an improvement of 6.3% compared to the traditional ST-GCN model, and the differentiation between similar sign language actions was significantly enhanced.

Real-time performance: The end-to-end interaction delay is less than 500ms, fully meeting the needs of real-time communication. The response time for hazard warnings is only 95ms, and the spatial positioning accuracy reaches 0.3 meters, providing safe and reliable environmental support for visually impaired users.

Lightweight and power consumption: The number of model parameters has been reduced by over 40%, with the device's operating power consumption stabilized at 1.2W. It supports continuous operation for over 8 hours, fully meeting daily usage needs.

Environmental adaptability: Under complex backgrounds and varying lighting conditions, the accuracy rate of sign language recognition fluctuates by no more than 3%. The accuracy rate of dangerous sound recognition reaches 92%, demonstrating strong resistance to environmental noise interference.

5.3 Result Analysis

The experimental results demonstrate that our system, through multimodal fusion technology and lightweight optimization scheme, successfully addresses the core pain points of traditional assistive devices: the binocular vision + SLAM solution achieves an integrated design of sign language 3D localization and environmental perception, effectively eliminating the physical constraints of traditional sensor gloves; the lightweight design of LT-STGCN + DSConv significantly

improves the system's real-time performance while ensuring recognition accuracy; the construction of a localized dataset effectively compensates for the lack of dialect sign language recognition. The stability and reliability of the system in complex scenarios fully verify the application value of multimodal perception technology in the field of barrier-free interaction.

6. Conclusion and Outlook

6.1 Research Summary

This study proposes and successfully implements an intelligent glasses barrier-free interaction system based on embedded multimodal perception. Through the collaborative work of a dynamic sign language recognition module, a multimodal tactile interaction engine, and a blind multimodal environmental perception module, an efficient and smooth bidirectional barrier-free communication bridge is constructed. The system adopts core technologies such as lightweight spatiotemporal graph convolutional networks, LSTM-CRF semantic mapping models, and improved VSLAM, achieving high accuracy, low latency, and low power consumption in multimodal information processing on an embedded platform. The dynamic sign language recognition accuracy reaches 87.6%, the end-to-end delay is less than 500ms, and the spatial positioning accuracy is 0.3 meters, fully meeting the practical needs of special groups.

6.2 Innovation Points

The Lightweight Spatio-Temporal Graph Convolutional Network (LT-STGCN) is proposed, innovatively combining the DSConv efficient convolution operator with the CoT Block temporal modeling enhancement technology to achieve embedded real-time recognition of complex continuous sign language.

Develop a multimodal haptic interaction engine that utilizes the LSTM-CRF model to facilitate context-aware conversion between speech, text, and haptics, supporting a wide range of semantic expressions.

Integrating an improved VSLAM and a dual-channel voiceprint recognition architecture, it achieves high-precision navigation and rapid hazard warning for visually impaired individuals, significantly enhancing environmental perception capabilities.

6.3 Future Outlook

Future research can be further expanded in three directions:

Continuously expanding the localized sign language dataset to cover more dialect variations and complex scene gestures, thereby further improving recognition accuracy;

Optimizing multimodal fusion algorithms and introducing the Transformer architecture to enhance cross-modal semantic association, thereby improving the processing capability of complex semantics;

Expanding functional application scenarios, such as incorporating a text-sign language synthesis module to support active communication from hearing individuals to deaf and mute individuals, achieving more comprehensive bidirectional interaction.

Acknowledgements

This work was supported by the "Jiangsu Provincial College Students' Innovation and Entrepreneurship Training Program (Project No.: S202513988012)".

References

[1] Meng Zhiqiang. Research on Human Action Recognition Method Based on Multimodal Graph Convolutional Neural Network. Changchun University of Technology, 2025

[2] Yang Li, Yin Shiqi, Wang Tingting. Object Detection in Optical Remote Sensing Images Based on YOLOv7. Infrared Technology, 2025, 47(11): 1398-1405

[3] Liu Xingzheng. Research on Gesture Recognition Algorithm Based on Improved YOLOv7. Anhui University of Science and Technology, 2023

[4] Zhao Zhiqi. Research on Visual SLAM Algorithms Based on Deep Learning. University of Electronic Science and Technology of China, 2025

[5] Wu Chundi. Research on Dynamic Sign Language Recognition Algorithm Based on Deep Learning. Shenyang University of Technology, 2024

[6] Tu Chong, Jin Liying, Wang Zhongren, et al. Overview of Speech Recognition Technology and Its Applications. Digital Technology and Application, 2025, 43(09): 179-181

[7] Wang Jingyao, Fan Fei, Liu Haoyu, et al. Deaf and Dumb Sign Language Recognition Based on Machine Vision - Voice Interaction System. Internet of Things Technology, 2021, 11(12): 3-5

[8] Song Bo. Research on Power Consumption Optimization of Intelligent Terminal Speech Recognition Chips. Electronic Components and Information Technology, 2025, 9(08): 31-33

[9] Fan Yuhan, Li Qianqian, Meng Xue, et al. Design and Implementation of a Gesture Recognition System Based on Binocular Vision Principle. Intelligent Internet of Things Technology, 2024, 56(06): 81-84

[10] Shi Zihao. Research on Semantic VSLAM Based on Deep Learning. Shenyang University of Technology, 2025

