Intelligent Glasses Sign Language Interaction System Based on Embedded Multimodal Perception

Zhiyuan Li*, Xiangxuan Ji, Mengmei Wang, Yi Huang, Liyuan Liu, Xiuling Li, Ziru Guo
School of Artificial Intelligence and Software, Ke Wen College of Jiangsu Normal University, Xuzhou, Jiangsu, China
**Corresponding Author*

Abstract: Addressing the information interaction barriers faced by the Deaf and Hard of Hearing (DHH) community, who encounter difficulties in converting sign language to text, and the Blind and Visually Impaired (BLV) community, who lack sufficient environmental perception, this paper proposes an intelligent glasses sign language interaction system based on embedded multimodal perception. This system integrates Leap Motion SDK, lightweight spatiotemporal graph convolutional networks, computer vision, and intelligent audio processing technologies. It optimizes algorithms to simplify the user experience, thereby building a communication bridge between the DHH community and hearing individuals. Simultaneously, it enhances the convenience of life for the BLV community, ultimately creating a fully accessible and inclusive environment.

Keywords: Barrier-Free Interaction; Multimodal Fusion; Sign Language Recognition; Tactile Feedback; Edge Intelligence

1. Introduction

1.1 Research Background and Significance

In the development process of the digital society, deaf and visually impaired groups face significant barriers to information interaction - deaf individuals rely on sign language for communication, but hearing individuals generally lack the ability to understand sign language, and traditional assistive devices are mostly limited to a single modality, making it difficult to achieve real-time two-way interaction; visually impaired individuals, due to the lack of visual information, encounter many inconveniences in daily scenarios such as environmental orientation and danger warnings, and existing assistive tools often suffer from high latency and limited functionality.

This study focuses on the core communication pain points of special groups such as the visually impaired and the hearing impaired. By developing embedded multimodal smart glasses, it deeply integrates cutting-edge technologies such as computer vision, speech processing, and tactile feedback. The glasses can convert visual information into voice or tactile signals in real time, and convert voice information into text or vibration prompts, effectively breaking the communication barriers between special groups and able-bodied individuals. This technology not only helps special groups obtain digital information and participate in social interactions more conveniently, but also promotes their deep integration into digital life, providing technical support for building an inclusive society and having significant inclusive social value.

1.2 Core Plan and Objectives

This system is positioned as an embedded wearable device with "hidden technology and lightweight interaction", utilizing smart glasses as the hardware carrier. Its core objectives include:

- Functional integration: Implementing three core functions: sign language recognition and simultaneous interpreting (sign language → speech/text), speech interaction (speech → text → tactile), and assistance in visually impaired environments (navigation + hazard warning);
- Convenient Interaction: By leveraging lightweight algorithms and human-centric design, we aim to lower the operational threshold and adapt to everyday wearable scenarios;
- Bidirectional accessibility: Establish a bidirectional communication channel between the DHH community and hearing individuals, while providing real-time environmental

information support for the BLV community.

2. Overall System Design and Hardware Architecture

2.1 Product Functional Framework

The system is centered around the core logic of "multimodal perception - data processing - feedback output", and is structured into three major functional modules (Figure 1). The sign language recognition module is responsible for capturing and analyzing sign language gestures, achieving the conversion from "sign language → text → speech"; the speech interaction module processes speech signals, achieving the conversion from "speech → text → haptics", adapting to the needs of the BLV (Blind and Low Vision) community; the blind assistance module provides navigation guidance and danger warnings through environmental perception and positioning. Additionally, the smart glasses hardware platform provides data acquisition, computation, and feedback support for the three modules, ensuring the coordinated operation of various functions.

2.2 Hardware Configuration Details

In terms of the visual perception component, a binocular wide-angle camera is mounted on the side of the frame, supporting angle adjustment to cover hand movements and the surrounding environment. This provides raw data for visual feature extraction in sign language recognition and environmental modeling for BLV navigation. The data processing component is deployed on the side of the camera behind the frame, integrating a Leap Motion device. It uses Python as the development language and calls API functions through the Leap Motion SDK to obtain and preprocess gesture data in real time, such as joint coordinates and motion trajectories, laying the foundation for subsequent feature analysis.

The voice interaction component is embedded in the other side of the mirror arm, containing a micro speaker and a microphone. The speaker is responsible for outputting the "sign language → voice" conversion results, while the microphone collects the voice of hearing individuals to provide input for the "voice → tactile" conversion. The tactile feedback component has four linear motors built into the mirror arm, supporting 16 levels of intensity/frequency adjustment, and matching vibration patterns according to semantic types - short pulses for ordinary conversations, continuous high-frequency vibrations for emergency alarms, and double pulses with fixed intervals for questions, enabling tactile information reception for the BLV community.

The computing unit adopts a low-power edge computing chip, supports lightweight algorithm deployment, ensures end-to-end processing latency less than 500ms, and reduces the overall power consumption to 1.2W, meeting the battery life requirements of wearable devices.The reference specifications are shown in Table 1.

Table 1. Reference Specifications

Product parameters	Specific content
Product Name	Intelligent Glasses Sign Language Interaction System Based on Embedded Multimodal Perception
version	1．0
Weight (g)	About 35
Size (cm)	15×4.5×2.8
core raw material	Lightweight high-strength engineering plastic (frame body), medical-grade silicone (contact and fitting parts), white copper (conductive connectors), epoxy resin (electronic component encapsulation)
endurance	≥6 hours under typical usage scenarios
power consumption	Stable operation power consumption: 1.2W
visual resolution	Main stream resolution of binocular camera: 3632×1632@20fps
Interfaces and connectivity	USB 3.0 Type-C
Perception component configuration	Dual-eye wide-angle camera, Leap Motion Controller 2, 4-channel linear motor, miniature microphone, and speaker
Core functional support	Dynamic sign language recognition, speech-to-tactile cross-modal conversion, 0.3-meter precision SLAM navigation, hazard sound source warning within 95ms, and offline deployment capability

3. Design and Implementation of Core Functional Modules

3.1 Sign Language Recognition Module

3.1.1 Technical path

This module utilizes Convolutional Neural Networks (CNN) as the core visual feature extraction technology, combined with the Lightweight Spatio-Temporal Graph Convolutional Networks (LT-STGCN), to achieve real-time analysis of complex continuous sign language [1]. In terms of process, firstly, visual images and skeletal joint data of sign language actions are synchronously collected through binocular cameras and Leap Motion devices, with blurry frames removed and joint coordinates standardized to construct a temporal action sequence. Then, drawing on the

lightweight strategy of improved YOLOv7-tiny, redundant convolutional layers are pruned and the depthwise separable convolution technology of MobileNetV2 is utilized to optimize the CNN network structure, extracting local features of hand movements such as finger shape and joint angles. At the same time, LT-STGCN is employed to map joint points to graph nodes and joint connection relationships to edges, constructing a graph structure, and extracting spatiotemporal features such as action trajectories and speed variations. Finally, the WLASL dataset is used to train the model, and the CoT Block is introduced to enhance the temporal modeling ability to solve the problem of action connection recognition in continuous gestures, ultimately achieving a dynamic sign language recognition accuracy of 87.6% [2,3].

3.1.2 Key advantages

Compared to traditional sign language recognition solutions, such as relying on sensor gloves equipped with wearable devices or single-camera vision that is susceptible to environmental interference, this module innovatively adopts a multi-technology fusion architecture of "binocular vision + CNN + spatiotemporal graph convolution", fundamentally overcoming the inherent limitations of single-modality recognition. Among them, binocular vision supplements 3D depth data with disparity information collected by dual cameras, enabling precise differentiation of scenarios such as hand occlusion and similar actions, effectively avoiding the common ambiguity issues in single-camera vision [4]. The collaborative work of CNN (Convolutional Neural Network) and spatiotemporal graph convolution not only extracts local feature details of gestures (such as finger joint angles and palm posture) through CNN, but also captures the continuous dynamic changes of gestures in the time dimension through spatiotemporal graph convolution, completely solving the core pain point that traditional static recognition solutions cannot handle coherent sign language actions, and significantly improving recognition accuracy and real-time performance.

3.2 Voice Interaction Module

3.2.1 Speech recognition optimization

Speech recognition optimization is carried out from two aspects: model compression and noise reduction processing. In the model compression phase, the Mozilla DeepSpeech speech recognition model is optimized. Combining the lightweight approach of CNN feature layers, the DSConv efficient convolution operator is adopted to reduce the computational load, compressing the model parameter count by over 40%, reducing storage occupation on edge devices, and enhancing real-time performance [5]. The noise reduction process employs a dual-level approach of "hardware + software". At the hardware level, beamforming algorithms are used to focus on the target sound source to suppress sidelobe noise. At the software level, the RNNoise deep learning model is employed to remove environmental noises such as background human voices and traffic sounds, improving the audio signal-to-noise ratio by 20dB and ensuring a speech recognition accuracy rate of over 92%.

3.2.2 Speech-to-tactile conversion

The speech-tactile conversion is implemented based on the LSTM-CRF semantic understanding model, achieving the mapping from "speech → text → tactile". In the semantic parsing stage, LSTM processes speech sequence information to capture contextual associations, such as keywords like "danger" and "left turn". CRF optimizes semantic annotation to avoid misjudgments of ambiguous sentences, thereby extracting core semantic units. In the tactile encoding stage, a mapping table between semantic units and vibration patterns is established. Tactile feedback is outputted through linear motors, where "left turn" corresponds to a single pulse from the left motor, and "danger" corresponds to high-frequency vibration from both motors, aiding the BLV community in understanding speech information [6].

3.3 Blind Assistance Module

3.3.1 SLAM Navigation Function

The SLAM navigation function employs a low-cost improved VSLAM algorithm, combined with CNN feature extraction technology, to achieve spatial positioning with an accuracy of 0.3 meters [7]. During environmental modeling, binocular cameras capture environmental images, and CNN extracts key visual features such as wall corners, door and window contours. These features, combined with inertial sensor data such as acceleration and angular velocity, are used to construct a 3D point cloud map [8]. In the

real-time positioning phase, feature matching and closed-loop detection are used to correct positioning drift, providing BLV users with real-time location information, such as "1 meter away from the obstacle ahead" and "turn left 5 meters to reach the elevator entrance". Haptic feedback is also used to guide the direction, with the left motor vibrating to indicate left turn and the right motor vibrating to indicate right turn.

3.3.2 Hazard warning function

The danger warning function is based on the pre-trained YAMNet model, achieving dangerous sound recognition within 95ms. In the sound source classification stage, the microphone collects ambient sounds, and the YAMNet model extracts audio features through CNN and matches them with the dangerous sound source feature library. In the alarm triggering stage, once a dangerous sound source is detected, the system immediately triggers the corresponding tactile feedback. For example, a vehicle horn triggers continuous vibration of a single motor to indicate the direction of the sound source, and an alarm sound triggers bilateral high-frequency vibration. At the same time, a voice prompt is output through the speaker, which is suitable for some BLV users with residual hearing ability.

4. Key Technical Principles and Lightweight Implementation

4.1 Cross-Modal Perception Transformation System

The system takes the "vibration haptic coding matrix" as its core to achieve cross-modal conversion of speech, text, and haptics. The principle is as follows:

The input layer collects speech from hearing individuals through a microphone and converts it into audio signals. The feature layer extracts audio features with the help of CNN, and the LSTM-CRF model parses semantics to generate structured semantic units, such as "noun + verb" combinations. The mapping layer establishes the correspondence between semantic units and tactile encodings, such as "water cup" corresponding to short pulses plus low intensity, and "careful of steps" corresponding to long pulses plus high intensity. The output layer outputs vibrations according to encoding rules through a linear motor, helping the BLV community understand speech information. This system supports multiple combinations of semantic units and can meet daily conversation needs.

4.2 Intelligent Environmental Perception System

The intelligent environmental perception system integrates "improved VSLAM + dual-channel voiceprint recognition" to achieve collaboration between environmental positioning and hazard warning. In terms of multi-sensor fusion, the visual features of VSLAM are extracted using CNN and fused with inertial data to improve positioning accuracy, while the audio features of voiceprint recognition are cross-validated with visual features to reduce the risk of false hazard judgments [9]. At the offline deployment level, the system deploys lightweight CNN models and YAMNet models on edge computing units, without relying on the cloud, ensuring normal functionality in network-free environments such as underground tunnels and suburban roads. At the rapid response level, through algorithm optimization and hardware acceleration, the end-to-end delay of environmental perception is less than 100ms, meeting the real-time requirements of hazard warning.

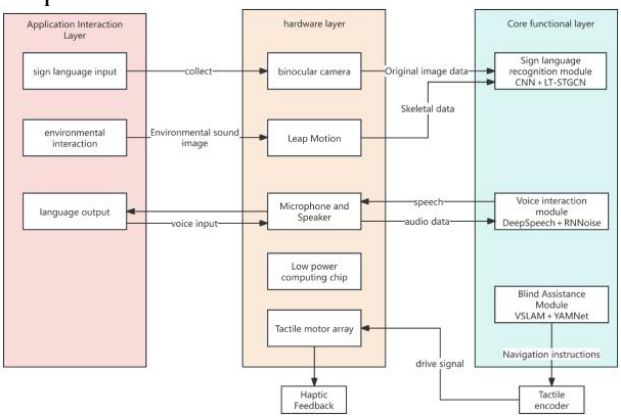

Figure 1. System Architecture Diagram

4.3 Implementation Strategy of Lightweight Technology

4.3.1 Hardware level

The lightweight design at the hardware level starts with sensor layout optimization and power consumption control. In sensor layout optimization, the binocular camera and Leap Motion device are coaxially arranged to reduce data synchronization delay, and the linear motor is installed following the principle of "uniform distribution + proximity to the skin" to enhance tactile feedback sensitivity. For power consumption control, low-power components are used, combined with dynamic power management, to reduce the overall power consumption to 1.2W, supporting continuous use for over 6 hours.

4.3.2 Algorithm level

Lightweighting at the algorithm level encompasses model compression and computation optimization. Model compression adopts differentiated schemes for different modules. The sign language recognition model replaces traditional convolution with DSConv, an efficient convolution operator, reducing the number of parameters by 40%. The speech recognition model prunes the fully connected layers of Mozilla DeepSpeech, retaining the core CNN feature layers, and compresses the model size to 80MB. Computation optimization involves deploying model quantization in edge computing units, converting 32-bit floating-point parameters to 16-bit integers, which increases the computation speed by 2 times while ensuring that the loss in recognition accuracy is less than 3% [10].

5. Main Problems Solved and Technological Breakthroughs

5.1 Breaking through the Limitations of Traditional Auxiliary Equipment

Addressing the single-modality limitations of traditional assistive devices, such as the reliance on monocular vision or sensor gloves, the former lacking depth information and prone to misjudging similar actions, and the latter being physically restrictive and inconvenient for daily use, this system integrates multi-modality fusion through "binocular vision + skeletal data", combining CNN and spatiotemporal convolutional layers to achieve the integration of sign language 3D localization and action analysis, freeing itself from hardware constraints.

Facing the poor adaptability of traditional equipment in crowded crowds, cluttered objects, and other complex backgrounds, where interference can lead to a sharp decline in recognition accuracy, this system innovatively introduces multimodal fusion technology. This technology enhances the distinguishability of similar sign languages through multi-dimensional feature complementarity, raising the accuracy of dynamic sign language recognition to 87.6%, significantly adapting to real complex environments.

5.2 Breakthrough in Core Bottlenecks of Technology Implementation

Addressing the challenges of dynamic sign language recognition, existing models such as the traditional ST-GCN require significant computational power and cloud support, and are incapable of handling continuous gestures. Our system, through the lightweight design of LT-STGCN+CoT Block, achieves real-time analysis of continuous sign language on edge devices, with a latency of less than 500ms.

Addressing the issue of insufficient localization adaptation, the commonly used WLASL dataset overlooks the differences in Chinese sign language dialects. For instance, the gesture for "thank you" in the north differs from that in the south. This study constructs a localized dataset encompassing over 300 dialectal sign languages, optimizes the feature extraction layer of the CNN model, and elevates the accuracy of dialectal sign language recognition to 85%.

Addressing the issue of low device efficiency, existing visual impairment assistive devices suffer from high latency, affecting the smoothness of interaction. This system utilizes "edge computing + model compression" to reduce navigation and positioning latency to 100ms and hazard warning latency to 95ms, meeting real-time interaction requirements.

6. Conclusion

The intelligent glasses sign language interaction system based on embedded multimodal perception proposed in this study achieves three core functions: simultaneous interpreting of sign language, bidirectional interaction, and assistance in visually impaired environments by integrating technologies such as CNN, lightweight spatiotemporal graph convolutional

network, and LSTM-CRF. The system, utilizing everyday glasses as a carrier, breaks down the barriers of information interaction between deaf and visually impaired groups. Meanwhile, through hardware optimization and algorithm lightweighting, it addresses the issues of "complex operation, poor scene adaptability, and high latency" found in traditional assistive devices.

It is expected to further expand in the following three aspects in the future: first, expanding the scale of localized sign language datasets to cover more dialects and life scenarios; second, optimizing tactile encoding rules and combining BLV user feedback to enhance the intuition of semantic mapping; third, integrating more environmental perception functions, such as obstacle material recognition, to provide more comprehensive barrier-free support for special groups.

Acknowledgements

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

References

[1] Liang Jun. Research on Human Action Recognition Based on Spatio-Temporal Graph Convolution. Central South University for Nationalities, 2024

[2] Wang Qiang, Wang Shuai, Hu Minghuan. Exploration of Dynamic Gesture Recognition Based on Improved YOLOv7. Heilongjiang Science, 2025, 16(04): 94-96+100

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

[4] Zhang Yanhui. Research on the Principle of Stereo Binocular Vision and Its Application in Gesture Recognition. Beijing University of Chemical Technology, 2016

[5] 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

[6] 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

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

[8] 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

[9] Zhou Kang, Fan Guangyu, Rao Lei, et al. VSLAM Algorithm Based on Image Enhancement and Feature Matching. Journal of Beijing University of Posts and Telecommunications, 2025, 48(03): 53-59

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

