

AI-Powered Multimodal Sign Language Recognition, Translation and Assistive Communication System

Abhisha KJ, Bachelor of Technology in CSE, NCERC

Akash KS, Bachelor of Technology in CSE, NCERC

Arathi M, Bachelor of Technology in CSE, NCERC

Ashik M, Bachelor of Technology in CSE, NCERC

Swathy CS, Assistant Professor, Department of CSE, NCERC

Abstract- Sign language is an essential communication medium for individuals with hearing and speech impairments, yet communication barriers still exist due to the limited availability of intelligent translation tools. This paper proposes an AI-powered multimodal sign language recognition, translation, and assistive communication system designed to improve accessibility and interaction between sign language users and the general community. The system utilizes MediaPipe to detect and extract 21 hand landmarks from sign gestures captured through a standard webcam. These landmarks are converted into structured JSON feature vectors and used to train a 1D Convolutional Neural Network (1D-CNN) model for accurate recognition of predefined as well as custom user-defined signs. The proposed system supports real-time and offline sign detection, sentence formation, multi-language translation, and text-to-speech output for enhanced communication. In addition, the platform enables multimodal interaction through voice, text, image, and video inputs. For video inputs, audio is extracted and converted to text, translated if necessary, and synchronized subtitles are embedded into the video. The system also includes a motivational support module that allows trained users to provide assistance and guidance. The proposed approach offers an inclusive, scalable, and hardware-independent solution for assistive communication.

Key Words: Sign Language Recognition, Deep Learning, MediaPipe, Multimodal Interaction, 1D Convolutional Neural Network

1. INTRODUCTION

Communication is an essential part of human interaction, but individuals with hearing and speech impairments often face difficulties when communicating with people who do not understand sign language. Sign language is widely used by deaf and hard-of-hearing individuals, yet the lack of accessible tools for real-time recognition and translation creates communication barriers in daily life.

Traditional sign language learning methods such as classroom training, printed charts, and prerecorded videos provide limited interaction and do not offer real-time feedback. Some existing technological solutions rely on specialized hardware like sensor gloves or motion tracking devices, which are expensive and inconvenient for regular use.

Recent advancements in artificial intelligence, computer vision, and deep learning have enabled the development of vision-based gesture recognition systems using standard cameras. In this paper, an AI-powered multimodal sign language recognition and assistive communication system is proposed. The system uses MediaPipe for hand landmark detection and a 1D Convolutional Neural Network (1D-CNN) model to recognize sign gestures. It supports real-time sign detection, sentence formation, multi-language translation, and text-to-speech output. The system also provides multimodal interaction through voice, text, image, and video inputs, making communication more accessible and inclusive for differently-abled individuals.

2. LITERATURE REVIEW

2.1 Hand Gesture Recognition using Computer Vision

Wang, Y., Liu, J., and Chen, X

Proposed a vision-based hand gesture recognition system using computer vision techniques for sign language interpretation. Their approach captures hand images using a camera and applies image processing methods such as segmentation and feature extraction to detect gesture patterns. Machine learning algorithms are then used to classify the gestures into predefined sign language symbols. The study demonstrated that vision-based approaches can significantly improve sign language recognition accuracy; however, the system performance was affected by lighting conditions and background variations.

2.2 Speech-to-Text Conversion using Google API

Zhang, Q. and Li, H.

Developed a speech recognition system using the Google Speech Recognition API to convert spoken language into text. The system processes audio signals and uses deep learning-based acoustic models to generate textual output. This approach enables deaf individuals to understand spoken communication through real-time text conversion. Although the system achieved good recognition accuracy, it faced challenges in noisy environments and with different accents.

2.3. Machine Learning-Based Gesture Classification

Kim, T., Park, S., and Lee, J.

Presented a gesture recognition system using machine learning algorithms such as Support Vector Machines and Convolutional Neural Networks. Their research focused on extracting gesture features from image frames and training classification models to recognize different hand gestures. The experimental results showed improved classification accuracy compared to traditional rule-based methods. However, the system required a large dataset for training to achieve reliable performance.

2.4 Text-to-Speech Systems for Accessibility

Taylor, P. and Black, A.

Conducted research on Text-to-Speech (TTS) systems designed for assistive technologies. Their work focused on converting textual information into natural and intelligible speech using natural language processing and deep neural networks. The proposed system improved speech quality and intelligibility, making it useful for communication support applications. Despite these improvements, early TTS systems had limited multilingual capabilities.

2.5 Neural Sign Language Translation

Camgoz, N., Hadfield, S., Koller, O., and Bowden, R. introduced a neural network-based approach for sign language translation. Their work used deep learning techniques to translate sign language videos directly into spoken language sentences. The proposed model combined convolutional neural networks and sequence learning methods to capture both spatial and temporal features of gestures. The study demonstrated the effectiveness of deep learning for automatic sign language translation but highlighted the need for large annotated datasets.

3. DESIGN METHODOLOGY

The proposed system follows a structured workflow to recognize sign language gestures, translate them into text or speech, and provide assistive communication through multimodal interaction. The system integrates computer vision, deep learning, and natural language processing techniques to ensure accurate gesture recognition and efficient communication support.

3.1 System Flow Diagram

The complete working model of the system is represented in the flow diagram below.

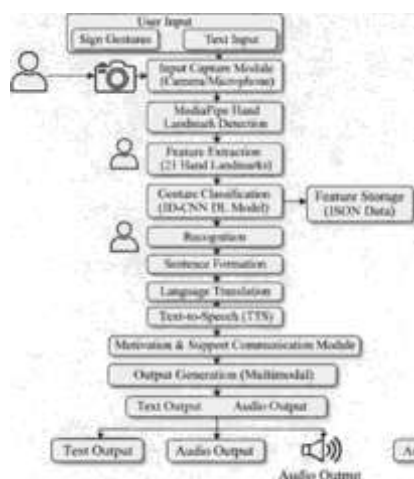


Fig 3.1.1: Flow Diagram

The system flow architecture describes the sequence of processes involved in converting user inputs into meaningful outputs. The overall workflow begins with user interaction and proceeds through several stages including hand landmark detection, feature extraction, gesture classification, sentence construction, translation, and output generation.

3.2 User Input

The system accepts multiple types of input such as sign language gestures through a webcam, voice input, text input, image input, or video input. For sign detection, real-time frames are captured using the device camera.

3.3 Hand Landmark Detection

The captured frames are processed using MediaPipe to detect the presence of hands and extract 21 hand landmark points. These landmarks represent the position of fingers and palm in three-dimensional space.

3.4 Feature Extraction

The detected landmark coordinates are converted into numerical feature vectors representing the gesture structure. These vectors are stored in a structured *JSON format*, which allows efficient dataset storage and expansion.

3.5 Sign Classification

The extracted feature vectors are provided as input to a 1D Convolutional Neural Network (1D-CNN) model. The CNN analyzes spatial patterns in the landmark data and predicts the corresponding sign gesture from the trained dataset.

3.6 Sentence Formation

The predicted signs are processed sequentially to remove repeated or incorrect detections. The filtered results are combined to form meaningful words or sentences.

3.7 Language Translation

The generated sentences are translated into the user's preferred language using natural language processing-based translation models.

3.8 Audio Output Generation

The translated text is converted into speech using a Text-to-Speech (TTS) engine, enabling audio communication for hearing users.

3.9 Video Subtitle Generation

For video inputs, the system extracts audio, converts it into text using speech-to-text technology, translates it if required, and embeds synchronized subtitles into the original video.

3.10 Output Delivery

The final output is delivered to the user in multiple formats such as text, speech, or subtitled video, ensuring accessibility and effective communication.

4. RESULT AND DISCUSSION



fig 4.1. User Module

The login page is designed with a clean and user-friendly layout to ensure easy access and secure authentication for users. At the center of the interface, a login panel is provided with input fields for Username and Password, allowing registered users to enter their credentials. After entering the required information, users can click the “Authorize Login” button to submit their details for verification.

The system validates the entered credentials against the records stored in the database. If the authentication is successful, the user is granted access to the system and redirected to the main application modules. This authentication mechanism helps prevent unauthorized access and protects the system's resources and user data. The interface also includes navigation options such as Home and Register, enabling new users to create an account and existing users to navigate easily.

The login interface is designed with a modern visual theme and responsive layout to enhance usability across different devices. Overall, this module acts as the first layer of security in the Nexus AI system, ensuring that only authorized users can access the sign language learning environment and its associated features.



fig 4.2. visual dictionary for learning

This interface is designed to help users understand and practice sign language through clear visual representations of hand gestures. The module displays alphabet cards arranged in a structured grid layout, where each card represents a specific alphabet letter along with its corresponding sign gesture.

Each card includes the alphabet label (such as A, B, C, and D) and an image showing the correct hand gesture used to represent that letter in Indian Sign Language. These images provide a visual guide that allows learners to easily observe and replicate the hand movements. The card-based interface improves readability and makes navigation simple for users, enabling them to browse through different alphabet signs efficiently.



fig 4.3 Multimodal input and translation interface

This interface represents the multimodal input module of the proposed system, where users can interact with the platform using different communication methods. The interface allows users to provide input through text or voice using the speech-to-text module. Once the input is provided, the system processes it through the neural input hub and stores it for further processing. The translation module converts the original input text into another language using an integrated language translation engine. The translated output is displayed along with an audio playback option using the text-to-speech module. This feature improves accessibility for both hearing and speech-impaired users by allowing them to communicate effectively across different languages. The interface also stores interaction history, enabling users to track previously processed inputs and outputs. This module demonstrates how multimodal interaction can enhance communication flexibility within the proposed assistive system.

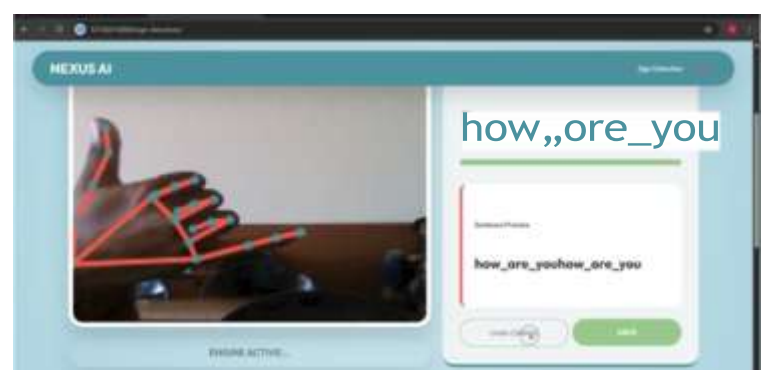


fig 4.4 Real-time sign language detection .

The sign detection interface demonstrates the real-time gesture recognition capability of the proposed system. In this module, the webcam captures hand gestures from the user and processes the input frames using the MediaPipe hand tracking framework. The system detects the hand region and identifies 21 landmark points representing the positions of fingers and palm joints. These landmark coordinates are used as input features for the trained 1D Convolutional Neural Network model, which classifies the detected gesture. The recognized gesture is then converted into text and displayed in the sentence preview section. The interface also includes options to edit, undo, or save the generated sentence. This functionality allows users to refine the detected text before storing it. The real-time visualization of hand landmarks helps users understand how the system interprets gestures and improves the transparency of the recognition process.



fig 4.5 Feedback htab interface for monitoring taser interactions and system performance.

The feedback hub interface allows users to provide feedback about the system's performance and interaction quality. This module is designed to improve the system continuously by collecting user responses and suggestions. Users can submit feedback by specifying the subject of the issue and describing their experience in the transmission details section. Once submitted, the feedback is stored and displayed in the interaction history panel, where the system tracks previous feedback entries and their processing status. This feature enables developers to monitor system performance, detect potential issues, and improve model accuracy or user interface functionality. The feedback module plays an important role in maintaining system reliability and enhancing the overall user experience by incorporating real user insights into the development process.



This interface is designed to assist users in converting spoken audio into readable text captions, thereby improving accessibility and comprehension, particularly for individuals with hearing impairments.

The interface provides a centralized panel titled "Upload Your Universe," where users can upload video files for subtitle generation. The system supports high-precision processing of video files up to 500 MB. Users can select a video by clicking the "Choose file" button and then initiate the process by pressing the "Ignite Subtitle Generation" button. Once the video is uploaded, the system begins processing the media using a multi-stage pipeline.

The workflow displayed at the top of the interface consists of four stages: Upload Media, English Transcription, Neural Translation and Final Burn-in. In the first stage, the video is uploaded to the server. The second stage converts speech in the video into text using speech recognition techniques. The third stage applies neural translation methods to refine the generated captions. Finally, the subtitles are embedded into the video during the burn-in stage.

The lower section of the interface displays the Processing Orbit, where users can monitor processing status, timestamps, and download the generated video or subtitle files (.SRT). This module enhances multimedia accessibility and supports inclusive communication through automated caption generation.

4. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my Mentor Ms.Swathy CS for their invaluable guidance, encouragement, and continuous support throughout the development of this project. Their expertise and insightful feedback played a crucial role in enhancing my understanding and improving the project's quality. I extend my heartfelt thanks to my team members for their dedication, teamwork, and commitment, which were instrumental in successfully implementing the AI-Powered Multimodal Sign Language Recognition, Translation and Assistive Communication System. Their contributions in various aspects of design, development, and testing greatly enriched the project. I am also thankful to Nehru College of Engineering and Research Centre, Thiruvillamala for providing the learning environment, enabling me to explore and apply my technical knowledge effectively. Finally, I deeply appreciate the unwavering support and encouragement from my family and friends, whose motivation kept me inspired throughout this journey. This project has been a significant learning experience, and I am truly grateful to everyone who contributed to its successful completion.

5. CONCLUSION

This paper presented an AI-powered multimodal sign language recognition and assistive communication system designed to reduce communication barriers for individuals with hearing and speech impairments. The proposed system integrates computer vision, deep learning, and natural language processing techniques to recognize sign language gestures and convert them into meaningful text and speech. By utilizing MediaPipe for hand landmark detection and a 1D Convolutional Neural Network (1D-CNN) model for gesture classification, the system achieves efficient and accurate sign recognition using a standard webcam without requiring specialized hardware. The platform also supports multimodal interaction through voice, text, image, and video inputs, enabling flexible communication methods for users. Additionally, features such as sentence formation, language translation, text-to-speech conversion, and video subtitle generation enhance the usability and accessibility of the system. The inclusion of a motivation and support module further strengthens the system by providing guidance and encouragement to users. Overall, the proposed system demonstrates the potential of AI-based assistive technologies in promoting inclusive communication and improving accessibility for differently-abled individuals.

6. REFERENCES

- [1] An, Yang, Baosong Yang, Binyuan Hui, et al. —Qwen2 Technical Report.11 arXiv preprint arXiv:2407.10671, 2024.
- [2] Athanasiou, Nikos, Mathis Petrovich, Michael J. Black, and Gül Varol. —SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation.11 In ICC V, 2023, pp. 9984-9995.
- [3] Baltatzis, Vasileios, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stelanos Zaleiriou. —Neural Sign Actors: A Diffusion Model for 3D Sign Language Production from Text.11 In CVPR, 2024, pp. 1985-1995.
- [4] Brown, Tom, Benjamin Mann, Nick Ryder, et al. —Language Models are Few-Shot Learners.11 NeurIPS, vol. 33, pp. 1877-1901, 2020.
- [5] Camgoz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. —Neural Sign Language Translation.11 In CVPR, 2018, pp. 7784-7793.
- [6] Devlin, Jacob, et al. —BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.11 arXiv preprint arXiv:1810.04805, 2018
- [7] Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, et al. —The LLaMA 3 Herd of Models. 11 arXiv preprint arXiv:2407.21783, 2024.
- [8] Huang, Wenca, Zhou Zhao, Jinzheng He, and Mingmin Zhang. —DualSign: Semi-Supervised Sign Language Production with Balanced Multi-Modal Multi-Task Dual Transformation.11 In ACM MM, 2022, pp. 5486-5495.
- [9] Hwang, Eui Jun, Huije Lee, and Jong C. Park. —Autoregressive Sign Language Production: A Gloss-Free Approach with Discrete Representations. 11 arXiv preprint arXiv:2309.12179, 2023.
- [10] Saunders, Ben, Necati Cihan Camgoz, and Richard Bowden. —Progressive Transformers for End-to-End Sign Language Production.11 In ECCV. Springer, 2020, pp. 687-705.