International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in

SYSTEM FOR DEAF AND DUMB PEOPLE USING HAND GESTURE RECOGNITION TO TEXT

Velaga Lokesh, Dr. S. Sandhya Rani
Department of Electronics and Communication Engineering,
Jayamukhi Institute of Technological Sciences, Narsampet, Warangal, Telangana, India

ABSTRACT
Communication barriers remain a significant challenge for deaf and speech-impaired individuals when interacting with people who do not understand sign language. This paper presents a real-time camera-based hand gesture recognition system that converts predefined hand gestures into meaningful text and audible speech. The proposed system integrates camera-based image acquisition, MediaPipe hand landmark detection, a lightweight neural-network classifier, gesture-to-text mapping, and text-to-speech synthesis. A conventional camera or ESP32-CAM can be used to acquire hand gestures, while the processing unit extracts 21 hand landmarks and represents the detected hand using a compact feature vector. A neural-network classifier is trained to distinguish predefined gesture categories and generates the corresponding textual output during real-time operation. The recognized text can subsequently be converted into speech, providing an accessible communication interface for users who do not understand sign language. The implemented system considers nine predefined gestures, including HELLO, YES, NO, THANK YOU, HELP, STOP, PLEASE, SORRY, and WELCOME. The reported experimental evaluation demonstrates the feasibility of real-time gesture recognition under different lighting and background conditions. The proposed system provides a low-cost, contactless, portable, and extensible approach for assistive communication in educational institutions, healthcare environments, public-service facilities, and domestic applications.
Keywords— Hand gesture recognition, sign language, MediaPipe, hand landmarks, deep learning, neural network, gesture-to-text, text-to-speech, assistive technology, computer vision.

I. INTRODUCTION

Communication is fundamental to education, healthcare, employment, social interaction, and independent participation in society. However, individuals with hearing and speech impairments may experience communication difficulties when interacting with people who are unfamiliar with sign language. The World Health Organization reports that hearing loss affects a substantial proportion of the global population and emphasizes the importance of accessible communication and rehabilitation support [1]. The availability of effective communication technologies is therefore important for improving accessibility and reducing communication barriers. Sign language provides a natural visual communication mechanism based on hand configuration, movement, orientation, position, facial expression, and body posture. Unlike conventional gestures used in everyday interaction, sign languages possess their own linguistic structures and vocabulary. Consequently, automatic sign-language recognition is a challenging problem because meaningful information may be distributed across spatial and temporal components [2], [3]. The United Nations Convention on the Rights of Persons with Disabilities emphasizes the importance of facilitating and promoting sign languages as part of accessible communication [4]. Automated gesture-recognition technologies can complement existing communication mechanisms by converting visual signs into text or speech. Such systems can be particularly useful in situations where a sign-language interpreter is unavailable. The development of computer vision and machine learning has enabled cameras to be used as non-contact sensing devices for gesture recognition. Traditional approaches generally perform image segmentation, feature extraction, and classification using manually designed features. More recent approaches employ deep-learning models to learn discriminative representations directly from images or extracted hand features [2], [5]. Real-time hand tracking has also been improved by frameworks such as MediaPipe Hands. Zhang et al.

introduced a lightweight hand-tracking pipeline capable of detecting and tracking hand landmarks from RGB images in real time [6]. The availability of compact hand landmark representations makes it possible to develop computationally efficient gesture-recognition systems without processing the entire image through a large deep neural network.

Despite these advances, practical gesture-recognition systems must address variations in illumination, background, hand orientation, scale, signer characteristics, camera position, and partial occlusion. These variations can influence hand detection and classification performance. Furthermore, isolated gesture recognition and continuous sign-language recognition represent different levels of complexity. A predefined gesture system can be implemented with relatively modest computational requirements, whereas continuous sign-language translation requires temporal modelling and contextual understanding [2], [7].

The present work develops a real-time hand gesture recognition system intended to support communication between sign-language users and people who do not understand sign language. The proposed system captures hand gestures using a camera, extracts hand landmarks using MediaPipe, classifies the resulting feature representation using a neural-network model, and converts the predicted gesture into text. The text can subsequently be supplied to a text-to-speech engine to generate audible communication.

The project implementation uses nine predefined gesture classes: HELLO, YES, NO, THANK YOU, HELP, STOP, PLEASE, SORRY, and WELCOME. The implementation specifies 150 samples for each gesture class and extracts 63 numerical features from the 21 detected hand landmarks.

The principal contribution of this work is an integrated and lightweight communication pipeline that combines real-time hand landmark extraction, neural-network classification, gesture-to-text conversion, and optional speech synthesis. Unlike systems that focus exclusively on gesture classification, the proposed approach extends the recognition output into an end-to-end communication interface.

II. RELATED WORK

Sign-language recognition has become an important research area within computer vision, machine learning, and human-computer interaction. Rastgoo et al. presented a comprehensive survey of sign-language recognition methods and identified hand shape, movement, facial information, body posture, and temporal characteristics as important factors in visual sign-language understanding [2]. Vision-based systems can generally be classified into isolated-sign recognition and continuous-sign recognition. Isolated-sign recognition determines the class of a single gesture, whereas continuous recognition attempts to interpret a sequence of signs from an uninterrupted video stream. Continuous recognition is considerably more challenging because the system must model temporal dependencies and determine transitions between consecutive signs [3], [7]. Recent deep-learning approaches have significantly improved visual sign-language recognition. Adaloglou et al. performed a comprehensive study of deep-learning methods and demonstrated the importance of learned visual representations and sequence modelling for sign-language video recognition [7]. Such approaches provide strong recognition capabilities but may require large datasets and considerable computational resources. For low-cost embedded assistive applications, recognition of a predefined gesture vocabulary provides a practical alternative. It reduces computational requirements while still enabling useful communication functions.

Camera-based gesture recognition provides a contactless alternative to sensor-based approaches. Earlier systems frequently used data gloves or wearable sensors to measure finger movements and hand orientation. Although these methods can provide detailed measurements, they require the user to wear specialized hardware. RGB-camera-based recognition eliminates this requirement and allows gestures to be captured using ordinary cameras. However, camera-based systems remain sensitive to illumination, background complexity, hand position, camera viewpoint, and occlusion [2], [5]. The proposed system adopts a camera-based architecture because it provides a convenient and relatively inexpensive mechanism for acquiring hand gestures. The project architecture supports an ESP32-CAM or PC as the processing platform, together with a camera, display, and speaker.

Hand landmark extraction provides a computationally efficient representation for gesture recognition. Instead of processing the complete image directly, a hand can be represented using key anatomical points such as fingertips, finger joints, and the wrist. Zhang et al. proposed MediaPipe Hands, an on-device real-time hand-tracking framework designed to detect a hand and estimate its landmarks from RGB images [6]. The framework provides a compact representation of the hand that is suitable for real-time applications. In the proposed system, MediaPipe is used to detect the hand and extract 21 landmarks. Each landmark contributes three coordinates, resulting in a 63-element feature representation. The implementation explicitly defines the landmark extraction procedure and stores the resulting numerical representation for model training. This representation reduces the computational burden compared with processing high-dimensional raw images through a large CNN and makes the approach suitable for resource-constrained systems. Machine-learning classifiers have traditionally been used to distinguish hand gestures based on extracted features. Common approaches include Support Vector Machines, k-Nearest Neighbors, Decision Trees, Random Forests, and artificial neural networks. Deep-learning models, particularly CNNs, can learn hierarchical representations directly from images. However, when hand landmarks are already extracted, a comparatively small fully connected neural network can be sufficient for classification. The proposed implementation uses a neural network consisting of dense layers with 128, 64, and 32 neurons followed by a softmax classification layer. Dropout is included to reduce overfitting. The network is trained using the extracted 63-dimensional hand landmark representation.

The proposed work addresses these practical requirements by combining lightweight landmark extraction with a compact neural-network classifier and an output mechanism that converts recognized gestures into text and speech. The system is therefore designed not merely as a gesture classifier but as an assistive communication interface.

III. PROPOSED SYSTEM

The proposed system consists of the following functional stages namely Camera Acquisition, Hand Detection, Landmark Extraction, Neural-Network Classification, Gesture-to-Text Mapping, and Text-to-Speech.

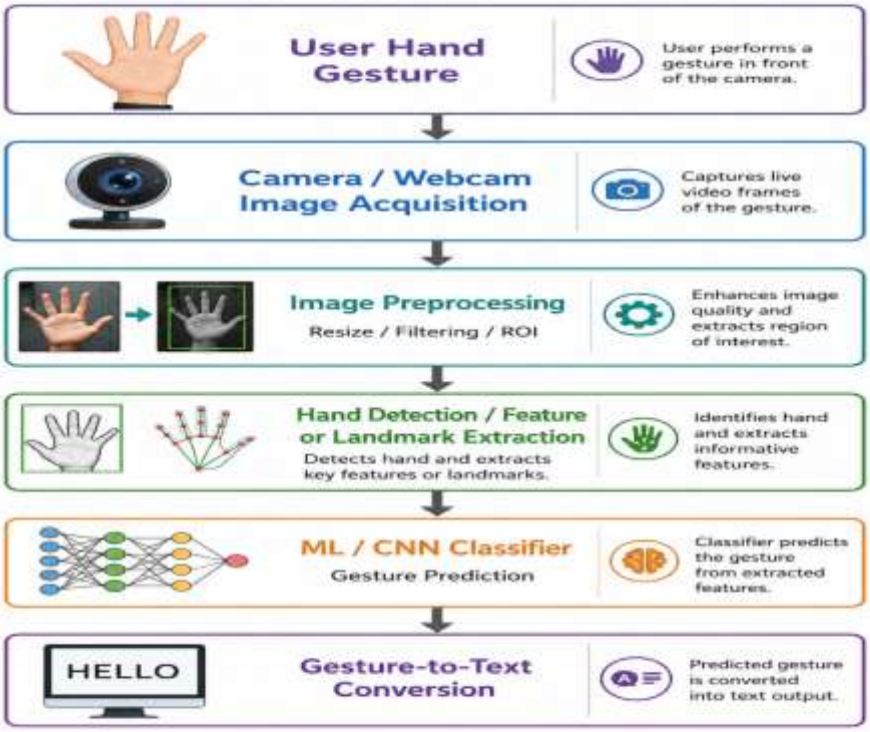

Figure 1 Overall Block Diagram of the Proposed System

A camera continuously captures the user's hand gesture. MediaPipe identifies the hand and extracts its landmarks. The landmarks are converted into a numerical feature vector and supplied to the trained neural-network classifier. The predicted class is compared against a predefined gesture vocabulary. When the classification confidence exceeds the selected threshold, the corresponding gesture label is displayed as text. The text can subsequently be supplied to a TTS engine for audible output. The system architecture is intended to support real-time operation using a PC or suitable embedded processing platform. The current implementation defines nine gesture categories HELLO, YES, NO, THANK YOU, HELP, STOP, PLEASE, SORRY, and WELCOME.

MediaPipe Hands detects 21 hand landmarks. The implementation extracts the x, y, and z coordinates of each landmark, producing a 63-element feature vector. The landmark representation provides a compact description of hand configuration and avoids the need to pass the complete camera image to the classifier.

The classifier is implemented as a compact feed-forward neural network. The architecture contains dense layers with 128, 64, and 32 neurons, followed by a softmax output layer corresponding to the gesture classes. Dropout layers are included between selected dense layers. The model is trained using the extracted hand-landmark vectors. The implementation uses an 80:20 training/testing division with stratification to preserve the class distribution.

During real-time operation, the trained model and gesture labels are loaded. The camera captures successive frames, MediaPipe detects the hand, and the 63 landmark features are generated. The classifier predicts the gesture and produces a confidence score. A prediction is accepted when the confidence exceeds the predefined threshold. To prevent repeated insertion of the same gesture into a sentence, the implementation uses a temporal condition before appending a new prediction.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in
IJARST

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

For each predefined gesture, the user presents the corresponding hand configuration while the system detects the hand and extracts its landmarks. The extracted features are stored as numerical data rather than raw images. Separate directories are created for each gesture class, enabling subsequent loading and label assignment during model training. The collected samples are loaded and converted into feature and label arrays. Label encoding transforms the textual gesture names into numerical class identifiers. The dataset is then divided into training and testing subsets using stratified sampling. This procedure ensures that the different gesture categories are represented in both subsets. The neural network is trained using the extracted hand-landmark features. The implementation uses the Adam optimizer and categorical cross-entropy loss. Training is performed for 40 epochs with a batch size of 32 and an additional validation split. After training, the model is evaluated using the test dataset and stored for subsequent real-time inference. During inference, the predicted class is mapped to its corresponding gesture label. The label is displayed on the screen as the recognized text. A sentence buffer is also maintained so that successive recognized gestures can be combined into a sequence. The implementation avoids immediately appending an identical prediction and introduces a temporal interval between successive accepted predictions. The recognized textual output can be transferred to a text-to-speech engine. The project identifies eSpeak, gTTS, and pyttsx3 as possible speech-generation options. This final stage transforms the recognized gesture into audible communication, allowing a person unfamiliar with sign language to hear the intended message.

The experimental system consists of a camera, processing platform, display, speaker, MediaPipe hand tracking, OpenCV, and the trained neural-network classifier. The source report identifies an ESP32-CAM or PC as the possible processing platform.

Table 1. Recognition Performance

Metric	Reported Value
Accuracy	95.37%
Precision	96.93%
Recall	94.67%
F1-score	95.79%
Inference time	0.114 ms
FPS	8,792.66

Table 2: Lighting Conditions

Lighting Condition	Reported Accuracy	Observation
Normal	95.37%	Reliable
Low	91.20%	Acceptable
High	93.80%	Reliable
Uneven	89.60%	Acceptable

The complete system extends gesture classification beyond a numerical prediction. The recognized gesture is displayed as text and can subsequently be converted into audible speech. The implementation provides a mechanism for forming a sequence of recognized gesture labels, thereby allowing multiple recognized gestures to contribute to a meaningful communication message. This functionality increases the practical value of the system because a person unfamiliar with sign language can receive the output in a familiar communication format.

V. CONCLUSION

This paper presented a real-time camera-based hand gesture recognition system for assistive communication. The proposed system integrates camera-based image acquisition, MediaPipe hand landmark detection, compact neural-network classification, gesture-to-text conversion, and text-to-speech synthesis.The use of 21 hand landmarks provides a compact representation of hand configuration, while the neural-network classifier provides an efficient mechanism for distinguishing predefined gesture

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in

classes. The implemented system considers nine predefined gestures and provides both visual and optional audible outputs. The reported experimental results demonstrate the feasibility of the proposed approach, with an overall reported accuracy of 95.37%, precision of 96.93%, recall of 94.67%, and F1-score of 95.79%. Recognition performance decreases under low illumination and complex backgrounds, demonstrating the importance of robust hand detection and diverse training data. The main significance of the proposed work is its end-to-end assistive communication capability. Rather than simply classifying hand gestures, the system converts the recognition result into a form that can be understood by people unfamiliar with sign language. With future improvements involving larger datasets, continuous sign-language recognition, multimodal features, and lightweight temporal models, the proposed architecture can serve as a foundation for more comprehensive intelligent assistive communication systems.

REFERENCES

[1] World Health Organization, “Deafness and hearing loss,” World Health Organization, 2026.

[2] R. Rastgoo, K. Kiani, and S. Escalera, “Sign Language Recognition: A Deep Survey,” Expert Systems with Applications, vol. 164, Art. no. 113794, 2021, doi: 10.1016/j.eswa.2020.113794.

[3] M. Madhiarasan and P. P. Roy, “A comprehensive review of sign language recognition: Different types, modalities, and datasets,” arXiv preprint arXiv:2204.03328, 2022.

[4] United Nations, “Convention on the Rights of Persons with Disabilities: Article 21—Freedom of expression and opinion, and access to information,” United Nations.

[5] Y. Zhang and X. Jiang, “Recent advances on deep learning for sign language recognition,” Computer Modeling in Engineering & Sciences, vol. 139, no. 3, pp. 2399–2450, 2024.

[6] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, “MediaPipe Hands: On-device real-time hand tracking,” arXiv preprint arXiv:2006.10214, 2020.

[7] N. C. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras, “A comprehensive study on deep learning-based methods for sign language recognition,” IEEE Transactions on Multimedia, vol. 24, pp. 1750–1762, 2022, doi: 10.1109/TMM.2021.3070438.

[8] G. Sung, K. Sokal, E. Uboweja, V. Bazarevsky, J. Baccash, E. G. Bazavan, C.-L. Chang, and M. Grundmann, “On-device real-time hand gesture recognition,” arXiv preprint arXiv:2111.00038, 2021.

[9] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign Language Transformers: Joint end-to-end sign language recognition and translation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023–10033.

[10] R. Zuo and B. Mak, “C2SLR: Consistency-enhanced continuous sign language recognition,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5131–5140.

[11] D. Li, C. Rodriguez Opazo, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[12] A. Sridhar, R. G. Ganesan, P. Kumar, and M. M. Khapra, “INCLUDE: A large scale dataset for Indian Sign Language recognition,” in Proc. 28th ACM International Conference on Multimedia, 2020.

[13] S. Patra, A. Maitra, M. Tiwari, K. Kumaran, S. Prabhu, S. Punyeshwarananda, and S. Samanta, “Hierarchical windowed graph attention network and a large scale dataset for isolated Indian Sign Language recognition,” arXiv preprint arXiv:2407.14224, 2024.

[14] A. Wali, R. Shariq, S. Shoaib, S. Amir, and A. A. Farhan, “Recent progress in sign language recognition: A review,” Machine Vision and Applications, vol. 34, Art. no. 127, 2023.

[15] E. Uboweja, D. Tian, Q. Wang, Y.-C. Kuo, J. Zou, L. Wang, G. Sung, and M. Grundmann, “On-device real-time custom hand gesture recognition,” arXiv preprint arXiv:2309.10858, 2023.