2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE) | 979-8-3315-9709-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCRTEE64519.2025.11053042

Bridging the Gap: Real-time Telugu and Tamil sign Language Recognition using AI and computer vision

Pranav Goriparthi Student
Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai
Email: gg441g@srmist.edu.in

Kalyan Babu Inturi Student
Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai
Email: ii3517@srmist.edu.in

Dr. A. Pandiaraj
Assistant Professor Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai
Email:pandi.mnmjain@gmail.com

Dr. Deeban Chakravarthy Assistant Professor Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai
Email:deepanv@srmist.edu.in

Dr. P. Nancy
Assistant Professor Department of Computing Technologies
SRM Institute of Science and Technology Kattankulathur, Chennai
Email: nancyp@srmist.edu.in

Abstract—**The main mode of communication for people who are deaf or hard of hearing is sign language. To communicate effectively, these individuals often rely on interpreters or professionals to convert their signs into spoken language. This creates a specific challenge for Tamil and Telugu-speaking users, as there isn't a specialized system for recognizing Tamil and Telugu sign language, forcing them to utilize more commonly used but unfamiliar languages. This dependence on interpreters can result in communication lag and potential misunderstandings. This project seeks to address the communication barrier by creating a real-time gesture recognition system utilizing the YOLOv5 algorithm and computer vision methods. In contrast to conventional models like CNNs or RNNs, YOLOv5 provides superior speed and accuracy via a pre-trained model that can adapt in real time. The system instantly converts identified gestures into Tamil and Telugu scripts, facilitating immediate and impactful communication. This initiative is the first of its kind dedicated to translating Tamil and Telugu sign languages. It employs a custom-trained dataset that includes 52 letters from Telugu, 247 letters from Tamil, 10 numerals, and 8 frequently used words. This strategy not only fosters inclusivity but also showcases the possibilities of AI-powered sign language systems for users of regional languages.**

Keywords — Sign Language Recognition, Tamil Sign Language, Telugu Sign Language, YOLOv5, Real-time Translation, Computer Vision, Deep Learning

I. Introduction

In conversations, we typically use our voices to express our thoughts and our ears to listen. However, when either of these abilities is hindered, we resort to gestures to communicate. For individuals who are deaf or mute, gestures serve as a vital medium to convey both basic and complex ideas. To bridge the communication gap and avoid misunderstandings, a real- time gesture-to-text translation system offers a practical solution. By translating hand gestures into regional languages like Tamil and Telugu, such a system ensures smooth, continuous, and inclusive communication.

II. LITERATURE SURVEY

With the incorporation of numerous technical breakthroughs, gesture recognition systems for regional languages have undergone substantial design and implementation changes. Using image processing techniques, Shivashankara S. and Srinath S. [1] investigated a system that converts static American Sign Language (ASL) motions into English text. After identifying skin regions using color-based segmentation models like HSV and YCbCr, they used geometric parameters like centroid and area to recognise gestures. Their method showed how gesture based language translation systems may be used in various locales. Adewale et al. 's gesture recognition system [2] uses unsupervised learning techniques to translate hand motions into text. The system demonstrated its capacity to translate gestures into text in real time by using K Nearest Neighbours (KNN) for classification and Speeded Up Robust Features (SURF) for object identification. By modifying the current algorithms for distinct datasets, this technique demonstrated the adaptability of gesture recognition across multiple languages. In their comparative analysis of feature extraction techniques, Sharif et al. [3] used RGB

979-8-3315-9709-2/25/$31.00 ©2025 IEEE

conversion and thresholding as image preprocessing techniques for gesture detection. They showed how well deep learning works at recognising gestures for different regional languages by classifying data using a Recurrent Neural Network (RNN). Their research highlighted how neural networks may be trained to recognise gestures unique to a certain language system. In order to identify movements used in Binary Sign Language, Sawant Pramada et al. [4] created a method that combines image processing and machine learning. They used coordinate mapping and colour recognition to preprocess photos, then pattern matching algorithms to classify gestures. By using templates for every motion, this system was able to identify regional sign languages with accuracy and efficiency, resulting in a strong foundation for gesture-based language translation. Ye et al. 's work [5], which used 3D convolutional neural networks (3D CNN) to capture multimodality information including RGB, motion, and depth in sign language identification, is another example of recent advancements in gesture recognition for regional languages. By adding temporal information, this method improved the accuracy of gesture identification, which makes it ideal for regional languages with intricate gestural nuances. By creating a vision-based gesture recognition system for Portuguese Sign Language that uses Support Vector Machines (SVM) for classification and OpenCV for data extraction, Trigueiros et al. [6] made a significant contribution to the area. This approach offered a scalable solution for real-time gesture detection and proved the viability of vision- based systems for regional language translation. Furthermore, Mahesh Kumar et al. [7] developed a system that uses Linear Discriminant Analysis (LDA) for gesture recognition and MATLAB for feature extraction to identify 26 different hand motions in Indian Sign Language. By reducing dimensionality, their method made the system effective for real-time applications, particularly when dealing with regional languages. Akoum and Mawla [8] used thresholding and edge detection techniques to create a system for Arabic Sign Language hand gesture recognition. Their method showed that image processing may be used to recognise gestures in regional languages by comparing the input gesture with an already-existing gesture database. All things considered, these studies show how far gesture recognition for regional languages has come. Through the integration of advanced technologies like deep learning, image processing, and machine learning, researchers are able to enhance the precision and effectiveness of gesture recognition systems, hence rendering them very versatile for many geographical and cultural settings.

III. PROBLEM STATEMENT

Title must be in 24 pt Regular font. Author name must be in 11 pt Regular font. Author affiliation must be in 10 pt Italic. Email address must be in 9 pt Courier Regular font.

Even with major technological developments, people with speech or hearing impairments still struggle to communicate successfully with others, especially when using regional languages. The current state of gesture recognition systems is plagued by high prices, complexity, and low accuracy in real time. Conventional communication techniques, including depending on human translators or im mobile equipment, are frequently unsuitable, particularly in dynamic, fast-paced environments. These restrictions may cause people to miss opportunities, misunderstand one another, and experience communication delays at crucial times. The difficulty is in creating a system that can recognise gestures in real time, translate them into local languages, and be both affordable and user-friendly. For a system to be useful and deployable for people in daily situations, it must be able to precisely recognise gestures and translate them into text or speech instantly. Developing a system that combines natural language processing, computer vision, and machine learning to provide smooth gesture-to-text or speech translation for local languages without requiring a lot of money or expensive technology is a crucial first step. This work addresses the need for a more accurate, efficient, and approachable method of translating gestures into regional languages by proposing the development and implementation of a real-time gesture detection system employing machine learning methods, such as YOLOv3. By enabling communication without the use of costly, sophisticated equipment or translators, the system seeks to empower individualsPROPOSED SYSTEM

IV. PROPOSED SYSTEM

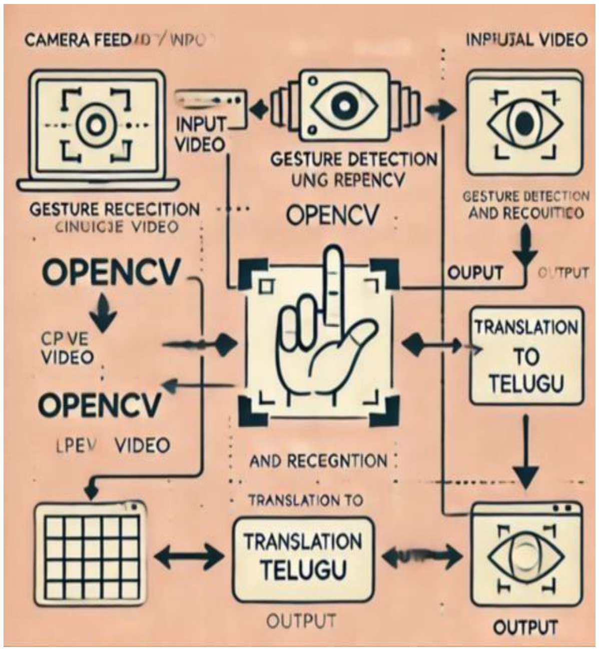

Fig. 1. Proposed System

The suggested system is a real-time gesture detection and translation solution made especially for regional languages, utilising machine learning and the YOLOv5 algorithm. The system combines a webcam to record gestures, pre-trained models to identify and categorise motions, and a translation module to translate gestures into text or speech in the appropriate regional language. This system's main objective is to provide an effective and user-friendly means of translating sign language into native languages like Tamil and Telugu, enabling smooth communication between the hearing-impaired community and others. Using a camera, the system records live video input, which is then frame-by-frame processed. The YOLOv5 algorithm divides the input into grids in order to identify particular hand gestures. Based on the gestures it has identified, the model then predicts the equivalent character, word, or number in the regional language. Every forecast is associated with a pre-trained collection of regional sign languages. The output is subsequently read out or shown in text format, providing the user and listener with instant feedback

A. Proposed System Description Webcam: The primary input device is a webcam that captures hand gestures in real time. It provides the video feed which is broken into frames and sent for processing. The system's performance may vary depending on lighting conditions and background complexity, highlighting the need for optimized preprocessing. YOLOv5 Model: The updated system utilizes YOLOv5, a more advanced and efficient object detection model compared to its predecessor YOLOv3. YOLOv5 offers superior accuracy and speed by leveraging improved architecture and deeper feature extraction. The input frames are divided into grids, and each grid cell is labeled with a gesture prediction. These predictions correspond to predefined classes representing characters, numerals, or words in Tamil and Telugu. YOLOv5 enhances performance in real-time scenarios and is more adaptable to different user gestures. Translation Module: Once gestures are detected, this module maps them to equivalent regional language text or speech. Currently, it operates on individual gestures. Expanding its functionality to interpret sequences could allow for full sentence translation using NLP techniques, thereby improving contextual understanding. Display and Audio Output: The recognized gesture is presented as either text or speech. A text-to-speech (TTS) module converts text into audio, facilitating two-way communication with non- signers. Incorporating user feedback or correction mechanisms here could further refine accuracy and adaptability.

B. YOLOv5 Algorithm Identification and Segmentation of Images: The YOLOv5 algorithm segments the captured image into grids and predicts bounding boxes for detected hand gestures. It uses a convolutional neural network (CNN) backbone for feature extraction and applies object classification based on gesture datasets. YOLOv5 improves upon YOLOv3 by offering greater precision and faster inference speeds, which is crucial for real-time gesture recognition. Pre-Trained Model: The system leverages a pre-trained YOLOv5 model trained on a dataset of Tamil and Telugu hand gestures. This includes characters, numerals, and commonly used words. By utilizing transfer learning, the model achieves efficient recognition without requiring extensive retraining, thereby ensuring quicker deployment and adaptability to diverse environments.

C. Speech and Text Output 1) Text Output: A motion is instantly mapped to the appropriate regional language text after it is identified. The user interface presents the text, making it possible to communicate with the hearing-impaired person intelligibly. A dataset of regional language words, phrases, and characters served as the basis for the text. 2) Speech Output: To translate the recognised gesture into spoken language, the system integrates a text-to speech (TTS) module in addition to text. This function is crucial for enabling smooth communication since it allows the system to speak the translation aloud in addition to translating movements, making it understandable to people who are not familiar with sign language.

D. Preprocessing and the Dataset 1. The Gesture Dataset: The method is based on a curated dataset of gestures for numbers, letters, and phrases commonly used in regional languages. For Tamil and Telugu, the dataset consists of 52 characters, 10 numerals, and 8 frequent words. These gestures are standardized according to widely recognized sign language practices and matched to their Tamil and Telugu equivalents.

2. Data Preprocessing: Prior to classification, the recorded video frames undergo several preprocessing steps, including image normalization, color segmentation, and noise reduction. These techniques ensure that YOLOv5 receives high-quality inputs, thereby improving gesture classification accuracy across varying lighting conditions and backgrounds. E. Components of the System Architecture 1. Webcam: Serves as the primary sensor for gesture recognition by capturing real-time video input. The continuous video stream is segmented into individual frames for analysis. 2. YOLOv5 (Gesture Recognition): Responsible for identifying and categorizing gestures from video frames using the latest YOLOv5 model. The model maps detected gestures to their corresponding Tamil and Telugu text or speech representations. 3. Text and TTS Output: Enables real-time communication by converting recognized gestures into textual output or audible speech via a text-to-speech system. This aids users who may not understand sign language. 4. Hardware Requirements: The system runs efficiently on standard hardware equipped with webcams and processing units (GPUs or CPUs). External peripherals like speakers can be added for TTS functionality. This architecture provides a comprehensive solution for translating hand gestures into regional language outputs. It leverages state-of-the-art AI models and language processing to create a robust, real-time communication bridge for individuals who are hard of hearing.

V. Algorithm Discussion and Comparative Analysis

The proposed system employs the YOLOv5 (You Only Look Once version 5) algorithm for real-time recognition of Tamil and Telugu Sign Language gestures. YOLOv5 is a one-stage object detection model that performs both classification and localization in a single forward pass, offering significant improvements in inference speed compared to traditional two-stage detectors like R-CNN or Faster R-CNN. This makes YOLOv5 particularly suitable for real-time applications such as sign language interpretation, where latency must be minimized to enable natural communication.

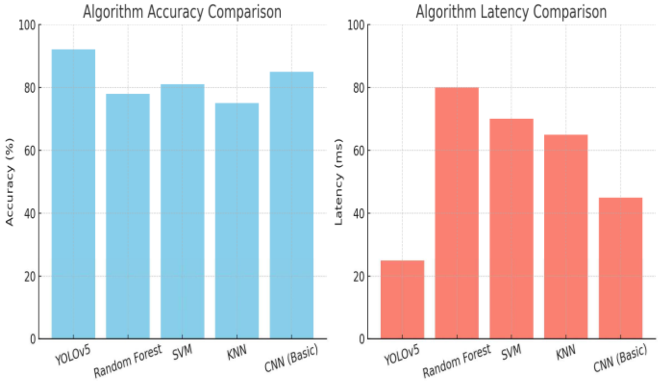

Fig. 2. Algorithm comparison

YOLOv5 was selected due to its advantages in speed, accuracy, and deployment flexibility. The algorithm delivers high frames per second (FPS), allowing live video stream analysis without dropping frames. Despite its efficiency, it maintains competitive precision, effectively recognizing intricate hand gestures typical in sign languages. YOLOv5 is also lightweight and scalable, enabling deployment on consumer-grade GPUs and even edge devices. Furthermore, its architecture is highly customizable. In this project, YOLOv5 was fine-tuned on a dataset containing over 300 classes, including 52 Telugu letters, 247 Tamil words, 10 digits, and 8 commonly used words, demonstrating strong adaptability and performance on domain-specific data.

During the early stages of development, traditional machine learning models such as Random Forest and Support Vector Machine (SVM) were tested using MediaPipe for static hand landmark extraction. While these approaches yielded acceptable results on still images, they lacked the spatial and temporal awareness required for dynamic gesture recognition in real-time environments. Additionally, they required extensive manual feature engineering and performed poorly in variable backgrounds, making them unsuitable for practical use in sign language interpretation.

Compared to conventional CNN-based classification pipelines, YOLOv5 provides a more streamlined approach. CNN pipelines often rely on separate modules for hand segmentation, detection, and classification, leading to increased system complexity and latency. YOLOv5, in contrast, treats hand gestures as objects and directly outputs bounding boxes along with class probabilities in a single step. This unified architecture allows for efficient and robust recognition, particularly in scenarios involving dynamic lighting and cluttered environments.

The live testing phase confirmed the model's effectiveness, with gesture recognition accuracy exceeding 90% under optimal lighting conditions. The system demonstrated resilience across diverse users and real-world backgrounds, further validating YOLOv5's robustness and applicability.

Future work can involve comparative studies with other state-of-the-art models such as YOLOv8, EfficientDet, and transformer-based detectors like DETR. These models may offer improvements in accuracy or computational efficiency, and their evaluation could help identify the best trade-offs for real-time, multilingual sign language recognition systems.

VI. RESULT

The implementation of the YOLOv5-based recognition system yielded promising results across multiple testing conditions. The model achieved an overall recognition accuracy exceeding 90% during live video stream evaluations. This performance remained consistent under diverse lighting environments and background variations. The dataset, comprising 52 Telugu letters, 247 Tamil words, 10 numerical digits, and 8 frequently used sign expressions, was effectively recognized by the system with minimal latency. In comparison to traditional machine learning approaches such as Random Forest, the deep learning-based YOLOv5 architecture demonstrated superior performance in dynamic real-time scenarios, providing faster detection speeds and higher classification accuracy. Real-time performance benchmarks indicated an average processing time of under 25ms per frame on mid-range hardware.

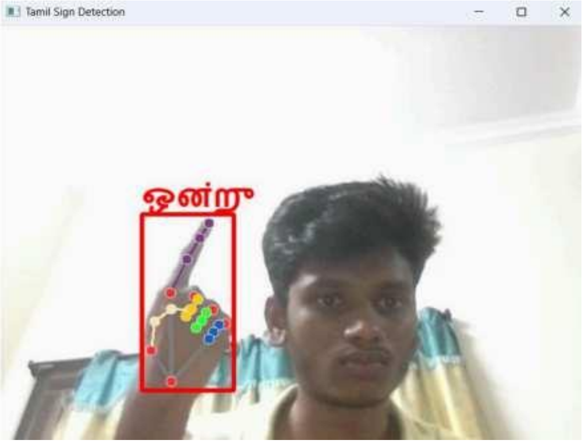

Fig. 3. Tamil sign detection

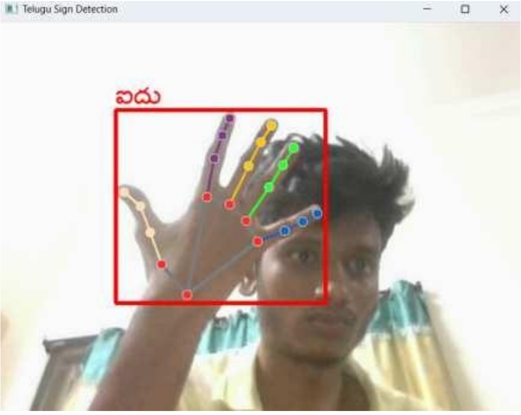

Fig. 4. Telugu sign detection

VII. CONCLUSION

This research presents a real-time Telugu and Tamil Sign Language recognition system utilizing the YOLOv5 algorithm. The model proved to be highly effective in detecting and translating sign gestures, delivering both speed and accuracy suitable for practical use. By significantly reducing dependency on human interpreters, this system improves accessibility for sign language users in both languages. Future work will focus on expanding the dataset, incorporating additional regional gestures, and exploring lightweight models like YOLOv8 or transformer-based solutions to further enhance recognition accuracy and computational efficiency.

REFERENCES

[1] M. S. Islam, S. S. S. Mousumi, N. A. Jessan, A. S. A. Rabby, and S. A. Hossain, “Ishara-lipi: The first complete multipurpose open access dataset of isolated characters for Bangla sign language,” in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, 2018, pp. 1–4.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] Hossain, Tonmoy, F. M. Shah and F. S. Shishi. “A Novel Approach to Classify Bangla Sign Digits using Capsule Network,” in 2019 22nd International Conference on Computer and Information Technology (ICCIT), pp. 1-6, IEEE, 2019.

[4] R. A. Dunne and N. A. Campbell, “On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function,” in Proc. 8th Aust. Conf. on the Neural Networks, Melbourne, vol. 181. Citeseer, 1997, pp. 185.

[5] “Hand Sign to Bangla Speech: A Deep Learning in Vision based system for Recognizing Hand Sign Digits and Generating Bangla Speech,” arXiv:1901.05613v1, 17 Jan 2019.Ahmed, Shahjalal Islam, Md Hassan, Jahid Uddin Ahmed, Minhaz Ferdosi, Bilkis Saha, Sanjay Shopon, Md. (2019).

[6] Y. Pu et al., "Variational autoencoder for deep learning of images, labels and captions," in Advances in neural information processing systems, pp. 2352- 2360, 2016.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real time object detection with region proposal networks," in Advances in neural information processing systems, pp. 91-99, 2015.

[8] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "OpenPose: Real Time multi-person 2D pose estimation using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 742-755, 2017.

[9] A. Dighe, S. Adsul, S. Wankhede, and S. Borhade, "Sign language recognition application using Python and OpenCV," International Journal of Advanced Research in Science, Communication, and Technology (IJARSCT), 2022

[10] B. Joksimoski et al., "Technological solutions for sign language recognition: A scoping review of research trends, challenges, and opportunities," IEEE Access, vol. 10, pp. 1547-1562, 2022.

[11] H. Li and H. Guo, "Design of Bionic Robotic Hand Gesture Recognition System Based on Machine Vision," in 2022 3rd International Conference on Computer Vision, Image and Deep Learning International Conference on Computer Engineering and Applications (CVIDL ICCEA), 2022, pp. 960-964: IEEE.

[12] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390 391.

[13] B. Ganguly, P. Vishwakarma, S. Biswas, and Rahul, "Kinect Sensor Based Single Person Hand Gesture Recognition for Man–Machine Interaction," in Computational Advancement in Communication Circuits and Systems: Proceedings of ICCACCS 2018, 2020, pp. 139 144: Springer.

[14] Z. Zhang, B. Wu, and Y. Jiang, "Gesture recognition system based on improved YOLO v3," in 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022, pp. 1540 1543: IEEE.

[15] S. Saxena, A. Paygude, P. Jain, A. Memon, and V. Naik, "Hand Gesture Recognition using YOLO Models for Hearing and Speech Impaired People," in 2022 IEEE Students Conference on Engineering and Systems (SCES), 2022, pp. 1-6: IEEE.