Indian Sign Language Recognition and Translation into Odia

Astha Swarupa Nayak[1], Naisargika Subudhi[2], Tannushree Rana[3],
Muktikanta Sahu[4], and Rakesh Chandra Balabantaray[5*]
Department of Computer Science and Engineering
International Institute of Information Technology Bhubaneswar
Odisha
[1]B121009@iiit-bh.ac.in, [2]B121031@iiit-bh.ac.in, [3]B421063@iiit-bh.ac.in,
[4]muktikanta@iiit-bh.ac.in, [5]rakesh@iiit-bh.ac.in

Abstract

Sign language is a vital means of communication for the deaf and hard-of-hearing community. However, translating Indian Sign Language (ISL) into regional languages like Odia remains a significant technological challenge due to the languages rich morphology, agglutinative grammar, and complex script. This work presents a real-time ISL recognition and translation system that converts hand gestures into Odia text, enhancing accessibility and promoting inclusive communication. The system leverages MediaPipe for real-time key-point detection and uses a custom-built dataset of 1,200 samples across 12 ISL gesture classes, captured under diverse Indian backgrounds and lighting conditions to ensure robustness. Both 2D and 3D Convolutional Neural Networks (CNNs) were explored, with the 2D CNN achieving superior performance 98.33% test accuracy compared to the 3D CNNs 78.33%. Recognized gestures are translated into Odia using a curated gesture-to-text mapping dictionary, seamlessly integrated into a lightweight Tkinter-based GUI. Unlike other resource-heavy systems, this model is optimized for deployment on low-resource devices, making it suitable for rural and educational contexts. Beyond translation, the system can function as an assistive learning tool for students and educators of ISL. This work demonstrates that combining culturally curated datasets with efficient AI models can bridge communication gaps and create regionally adapted, accessible technology for the deaf and mute community in India.

1 Introduction

Communication is a fundamental aspect of human interaction. However, individuals with hearing and speech impairments face significant barriers in engaging in verbal communication with the wider community. To overcome these challenges, sign language has been developed and adopted as an effective visual-gestural medium of communication. Sign language comprises hand gestures, movements, facial expressions, and body postures to convey meaning. Several physical and dynamic parameters, such as hand shape, hand orientation, motion trajectory, spatial positioning, and non-manual signals like facial expressions, play crucial roles in forming meaningful signs.

Globally, there are over 200 distinct sign languages, each with its own grammatical rules and syntactic structures (World Federation of the Deaf). Indian Sign Language (ISL) is one such rich and complex language that is widely used by the deaf community across India. ISL is not a direct manual representation of spoken Indian languages; rather, it possesses its own unique linguistic features and visual grammar. Although Indian Sign Language (ISL) serves as an effective communication medium for the deaf community across India, communication gaps still arise when regional languages-speakers are unfamiliar with ISL. In regions like Odisha, where Odia is the dominant spoken language, the absence of accessible translation systems between ISL and Odia limits seamless interaction. This communication barrier can hinder educational, social, and professional opportunities for hearing-impaired individuals when they engage with Odia-speaking communities.

The diversity of sign languages, along with regional variations within ISL itself, further complicates mutual understanding. Moreover, translating ISL into regional spoken languages like Odia, which is predominantly used in the Indian state of Odisha, can help bridge this gap and promote inclusivity. Developing a real-time ISL-to-Odia translation system has the potential to empower hearing-impaired individuals by enabling smoother interaction in academic, professional, and social environments.

[*] Corresponding author: rakesh@iiit-bh.ac.in

Proceedings of the Workshop on Sign Language Processing (WSLP), pages 10–17
December 20-24, 2025 ©2025 Association for Computational Linguistics

Sign language recognition systems primarily aim to track and interpret dynamic hand gestures and poses. However, building an accurate recognition system introduces several challenges. Vision-based sign recognition systems are prone to environmental factors such as varying lighting conditions, complex backgrounds, skin tone variations, and occlusions, which can hinder accurate gesture detection. To minimize these challenges and ensure robust gesture tracking, advanced computer vision frameworks like MediaPipe Holistic, developed by Google, can be leveraged. MediaPipe Holistic provides highly accurate real-time tracking of hand landmarks, pose, and facial key points, which are essential for extracting reliable sign features.

While several studies have successfully applied CNNs for sign language recognition in American Sign Language (ASL) (Natarajan et al., 2022) and British Sign Language (BSL), limited research exists for Indian Sign Language (ISL) and its translation into regional languages like Odia. Given the large hearing-impaired population in India and the cultural relevance of Odia, an ISL-to-Odia translation system is both necessary and impactful.

Communication is a fundamental human right that enables participation in educational, social, and professional spheres. However, individuals from the hearing and speech-impaired community often face significant communication barriers, especially in multilingual regions like India. Most existing technological solutions for sign language recognition and translation predominantly focus on translating ISL into English or Hindi. These systems overlook the linguistic diversity of India and fail to cater to regional languages such as Odia. Considering that approximately 18.9% of persons with disabilities in India reported hearing impairments (Census of India, 2011), and with over 42.5 million Odia speakers nationwide (Census of India, 2011), there is a significant need for accessible communication technologies tailored to this linguistic group.

In Odisha, the absence of a real-time ISL-to-Odia translation system poses a significant barrier for the deaf and hard-of-hearing community. Without accessible tools, effective communication with Odia-speaking peers, educators, and service providers remains limited, leading to social exclusion, reduced educational access, and restricted professional participation.

Real-time ISL-to-Odia translation is technically challenging due to the complexity of sign language, which involves dynamic hand gestures, facial expressions, and body movements. Accurate, real-time translation requires advanced computer vision and deep learning models capable of handling spatial and temporal features. Challenges also include the lack of comprehensive ISL-Odia datasets, difficulties in direct word mapping, and the need for lightweight models suitable for real-time use. To bridge this communication gap and promote inclusivity, there is a pressing need for a robust ISL-to-Odia translation system that can accurately recognize ISL gestures and generate grammatically correct Odia text in real time, enabling seamless interaction between hearing-impaired and hearing individuals.

This work aims to address this unmet need by developing a real-time ISL-to-Odia translator that leverages MediaPipe Holistic for landmark detection, a convolutional neural network (CNN) for gesture classification, and a custom ISL dataset. The system ensures accurate and culturally aligned translation, tailored to the linguistic and contextual nuances of the Odia language.

2 Literature Review

One of the initial approaches we studied was based on the VGG19 model (Shanavas et al., 2024), a deep convolutional neural network known for its high accuracy in image classification tasks. It could process hand gestures at 30 frames per second with an impressive accuracy of 95%, while maintaining robustness against lighting and background variations. However, its single-modal nature and focus on static image processing made it less effective in highly dynamic environments where sign gestures change rapidly over time, reducing its reliability for real-time applications.

Another promising technique utilized a combination of MediaPipe for extracting key facial, hand, and pose landmarks, and Long Short-Term Memory (LSTM) networks (Rehan and Mullick, 2023) for capturing temporal dependencies across gesture sequences. This method demonstrated improved understanding of dynamic sign gestures and achieved an accuracy range of 91% to 93%. Along with giving low accuracy, the system also relied on a fixed 30-frame input sequence, which limited its flexibility and responsiveness in real-time interactive settings.

A separate approach leveraged OpenPose

(Neyra-Gutiérrez and Shiguihara-Juárez, 2020) for keypoint detection and applied neural network-based summarization techniques for recognizing Peruvian Sign Language (PSL). It achieved an accuracy of 91.56% and was found to be computationally efficient. However, the model did not incorporate 3D keypoints or facial cues, resulting in reduced expressiveness and context-awareness in gesture interpretation. In another study, a hybrid model combining 3D Convolutional Neural Networks with Support Vector Machines (SVM) was employed to extract spatial and temporal features from gesture sequences in Chinese Sign Language (CSL) (Zhao et al., 2021). This method achieved 92.6% accuracy and showed strong capability in modeling motion patterns. However, its slower recognition speed and dependence on limited CSL datasets made it less suitable for real-time deployment and regional language adaptation.

In (Himasree et al., 2024), the authors proposed a novel Sparse Gabor Descriptor (SGD)-based technique along with random forest for gesture recognition with an accuracy of 94%. Similarly, The system utilizes a Vision Transformer (ViT) trained on a comprehensive video dataset to classify various sign language elements, while integrating a sophisticated language model, PHI-1.5B, to refine translated text for grammatical correctness and structural integrity and achieved robust and contextually relevant translation of ISL gestures into textual representations in (P and Francis, 2024).

The system proposed in (Kondo et al., 2024) employs the Mediapipe pose estimation library to pinpoint the exact positions of finger joints within video frames and converts these positions into one-dimensional angular features. These features are then organized sequentially to create a two-dimensional input vector for the ViT model.

The authors proposed a progressive sign language translation model to effectively separate sign language users from the background and reduce environmental interference, thus significantly improving the generalization ability in (Zou et al., 2024).

In another work proposed in (Prabha et al., 2024), the system focuses on breaking down video input into individual image frames and building three different models: EfficientNetV2, EfficientNetV2L, and ConvNeXtLarge algorithm. The accuracy yielded by the three models EfficientNet_V2, EfficientNet_V2L and ConvNextLarge are 94.20%, 92.54% and 95.21% respectively.

In order to identify an Isolated Sign Word (ISW) in Continuous Sign Language Videos (CSLV), aka Sign-Spotting, the authors proposed a Grammar-Based Inductive Learning (GBIL) framework utilizing a Grammar-Based Dictionary (GBD) that comprises pre-defined syntactic structures of tokens for handshape, location, and movement related to every Isolated Sign Word. GBIL can improve the cross-domain performance of sign spotting by integrating a grammar logic-based inference on top of deep learning architectures in (Amperayani et al., 2024).

The authors presented R-SLR, a sign language recognition system that can recognize the signs in real-time in (Ghosh et al., 2024). R-SLR identifies the hand in a video stream and extracts the region of interest. We extract the features from the preprocessed frames and classify the signs using the pre-trained DenseNet 201 model. The model performance is tested and it achieves 96.5% accuracy.

Recent research introduced an LSTM-based model with MediaPipe Holistic for Bangla Sign Language (BdSL) recognition (Das et al., 2025), achieving 88.33% accuracy by extracting keypoints and analyzing temporal gesture sequences. While effective for translating 100 isolated signs in real-time, the system faces limitations in vocabulary coverage, sentence formation, and sensitivity to lighting conditions, restricting broader usability.

Traditional gesture recognition algorithms (Badhe and Kulkarni, 2015) using FFT (Fast Fourier Transform) and template matching also showed promising accuracy (97.5%) for ISL. However, their rigid architecture, limited flexibility, and reliance on predefined gesture templates made them less adaptable for dynamic and continuous sign language input in real-world settings.

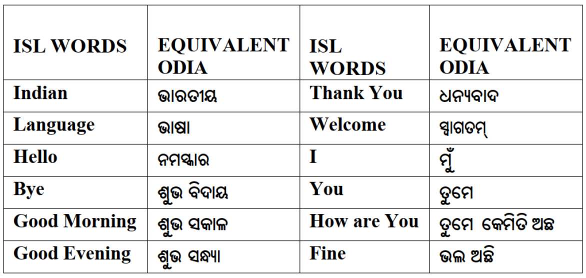
After evaluating all these models, the approach that stood out as the most relevant for our objectives was the CNN-based model tailored for Indian Sign Language and American Sign Language. This approach achieved the highest accuracy of 99.72% (Antad et al., 2024) among the surveyed models. It used convolutional neural networks for real-time detection of hand gestures, offering a practical blend of high performance, computational efficiency, and ease of implementation. The model's proven effectiveness with ISL and its adaptability to regional translation tasks made it an ideal choice for the current stage of our project.

Based on this comprehensive analysis, we concluded that the CNN-based model best met the requirements of our system. Its high accuracy, real-time processing capability, and compatibility with Indian Sign Language made it highly suitable for building our ISL recognition and translation system aimed at converting sign gestures into Odia text, thereby enhancing communication accessibility for the Odia-speaking deaf community.

3 Proposed Solution

To build a robust and context-aware gesture recognition system, we utilize OpenCV in combination with MediaPipe Holistic to capture human body landmarks in real time using a webcam. MediaPipe Holistic provides comprehensive tracking of facial features, body posture, and hand movements, which is essential for accurately detecting and interpreting sign language gestures.

The captured key pointscomprising coordinates of various body, face, and hand landmarks are extracted frame-by-frame and stored in structured files. These files form the basis of a custom dataset specifically designed for gesture recognition tasks. We focus on 12 commonly used ISL gestures: Indian, Language, Hello, Bye, Good Morning, Good Evening, Thank You, Welcome, I, You, How Are You, and Fine. Each gesture was recorded multiple times to capture variations in style, speed, and hand positioning. The final dataset consists of 1,200 samples (100 per class), collected under diverse Indian backgrounds and lighting conditions, and processed using Media Pipe for real-time key point extraction. Figure 1 shows the mapping of these commonly used gestures to Odia.



ISL WORDS	EQUIVALENT ODIA	ISL WORDS	EQUIVALENT ODIA
Indian	ଭାରତୀୟ	Thank You	ଧନ୍ୟବାଦ
Language	ଭାଷା	Welcome	ସ୍ୱାଗତମ୍
Hello	ନମସ୍କାର	I	ମୁଁ
Bye	ଶୁଭ ବିଦାୟ	You	ତୁମେ
Good Morning	ଶୁଭ ସକାଳ	How are You	ତୁମେ କେମିତି ଅଛ
Good Evening	ଶୁଭ ସନ୍ଧ୍ୟା	Fine	ଭଲ ଅଛି

Figure 1: Common Indian Sign Language (ISL) Words and Their Equivalent Odia Translations.

To make the model robust to real-world conditions, all gestures are captured in complex Indian backgrounds, featuring variations in lighting, background objects, and clothing. This ensures the dataset reflects real-life environmental complexity and improves the model's ability to generalize during real-time deployment.

This carefully curated and diverse dataset enables the training of a reliable sign language recognition model tailored for Indian cultural and visual contexts.

The proposed system presents a comprehensive and innovative approach to Indian Sign Language (ISL) recognition and its translation into the Odia language, with a strong focus on real-time applicability and inclusivity. It leverages the synergy between MediaPipe and Convolutional Neural Network (CNN) architectures for accurate gesture recognition, followed by dictionary-based mapping for regional language translation.

The integration of Mediapipe and CNN architecture within the system follows a cohesive and structured approach. Video frames captured by the webcam are processed using Mediapipe to extract relevant landmarks and features corresponding to facial expressions, body poses, and hand gestures. These extracted features are then fed into the CNN architecture for further analysis and classification, resulting in the recognition of specific sign language gestures. The overall system flow ensures seamless interaction between different components, enabling efficient and accurate sign language interpretation in real-time scenarios. By leveraging the capabilities of Mediapipe and CNN architecture, the system architecture demonstrates a powerful and effective approach to sign language recognition. Through continuous refinement and optimization, the system aims to provide enhanced support for individuals with hearing and speech impairments, empowering them to communicate effectively and participate fully in society.

To build a robust ISL-to-text translation system, a custom dataset was collected using a webcam and the MediaPipe library, capturing a diverse set of signs, including greetings, numbers, and alphabets. MediaPipe enables real-time extraction of normalized hand key-points, ensuring consistent input that is unaffected by background or lighting conditions. These key-points are preprocessed and fed into deep learning models for gesture classification. For model selection, both 2D CNN and 3D CNN architectures were implemented and evaluated to determine the one best suited to our performance and system requirements.

We developed and compared two custom multilayered models, one comprising 2D CNN layers and another comprising 3D CNN layers, to determine the optimal model for our sign language translation system. This comparative approach allowed us to identify the most computationally efficient architecture that maintains high accuracy for real-time translation of sign language gestures, balancing performance requirements with the temporal modelling capabilities essential for capturing sequential hand movements. The detailed rationale and architectural overview of the proposed system is elaborated in Figure 2.

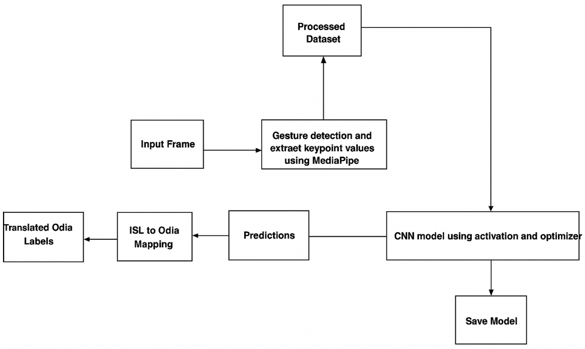

Processed Dataset
Input Frame
Gesture detection and extraet keypoint values using MediaPipe
Translated Odia Labels
ISL to Odia Mapping
Predictions
CNN model using activation and optimizer
Save Model

Figure 2: System Architectural Pipeline.

4 Results and Discussion

To assess the effectiveness of the proposed Indian Sign Language (ISL) to Odia text conversion system, we implemented and evaluated two distinct deep learning models: a 3D Convolutional Neural Network (3D CNN) and a 2D Convolutional Neural Network (2D CNN). Both were trained and tested on a custom ISL gesture dataset comprising 12 classes, captured in diverse Indian backgrounds to enhance real-world adaptability. The models were assessed based on training accuracy, validation accuracy, test accuracy, generalization, convergence speed, and deployment feasibility.

The 3D CNN was trained on short video sequences of 30 frames (84Œ20), allowing the model to learn spatiotemporal dynamics of gestures. The architecture consisted of stacked 3D convolutional blocks with Conv3D, Batch Normalization, MaxPooling3D, Dropout, and Dense layers. Despite its ability to capture temporal transitions, the 3D CNN demonstrated slower convergence and lower generalization:

- Training Accuracy: 88.62%
- Validation Accuracy: 77.63%
- Test Accuracy: 78.33%

Although the model improved over 75 training epochs, its validation and test accuracy (shown in Figure 3 and Figure 4) lagged, showing signs of overfitting. This outcome suggests that while 3D CNNs are suited for motion-aware tasks, the gesture variability and limited dataset size hinder their generalization. Additionally, its large parameter count (~570K) increased the risk of resource consumption.

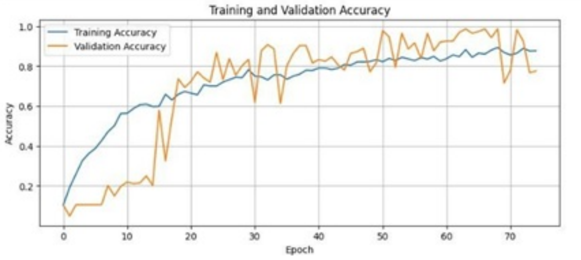


Figure 3: Training and Validation Accuracy of 3D-CNN.

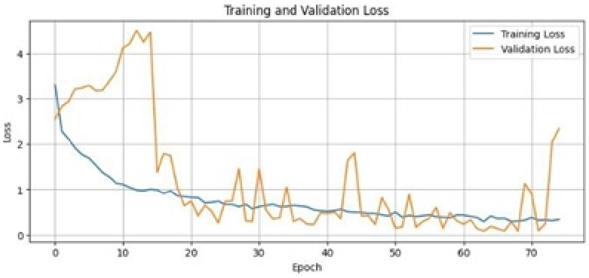


Figure 4: Training and Validation Loss of 3D-CNN.

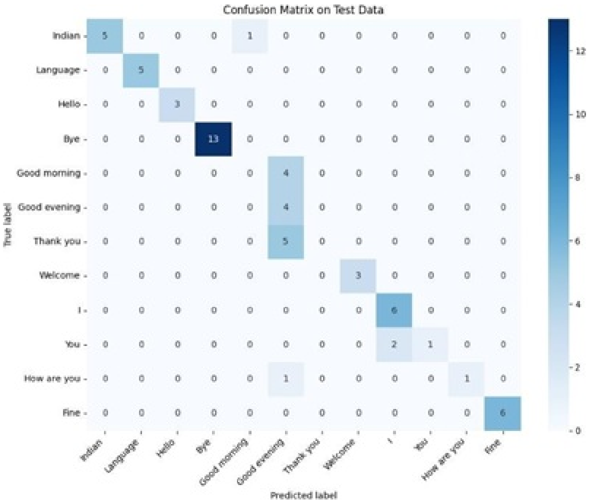


Figure 5: Confusion Matrix of 3D-CNN.

The 3D CNN confusion matrix (shown in Figure 5) reveals several misclassifications despite correct predictions in many categories. For instance, it confuses “Good evening” with “Good morning,” and one sample of “How are you” is misclassified, indicating challenges in capturing fine-grained spatial features despite temporal modelling. Key errors included confusing “Indian” with “Bye”, “Good evening” with “Good morning”, “You” with “I” and “How are you” with “Good evening”, along with two other isolated misclassifications. These results indicate that the model had difficulty distinguishing between gestures with subtle spatial or temporal similarities, leading to reduced accuracy in certain classes.

A more efficient 2D CNN was developed, utilizing skeletal keypoint features (x, y, z coordinates) extracted from each frame using MediaPipe Holistic. These features were flattened and treated as input for the model. The 2D CNN, built using Conv2D, Batch Normalization, MaxPooling2D, Dropout, and Dense layers, showed remarkable performance showing:

- Training Accuracy: 95.04%
- Validation Accuracy: 99.56%
- Test Accuracy: 98.33%

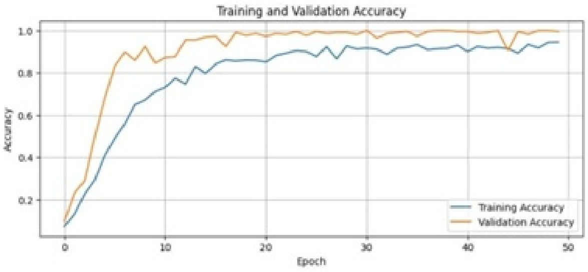


Figure 6: Training and Validation Accuracy of 2D-CNN.

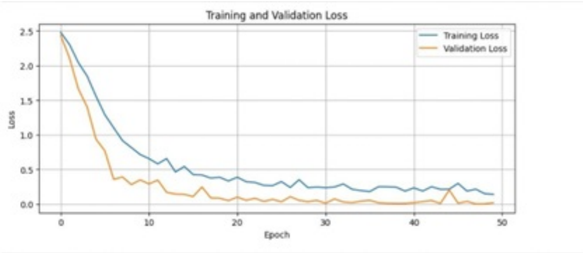


Figure 7: Training and Validation Loss of 2D-CNN.

Figure 6 and Figure 7 show the respective graphs relating to training and validation accuracy and loss obtained. The model not only converged faster (within 50 epochs) but also generalized better on unseen data. It was less prone to overfitting, required fewer computational resources (~160K parameters), and performed well under varying lighting, orientation, and hand shape conditions commonly found in Indian settings.

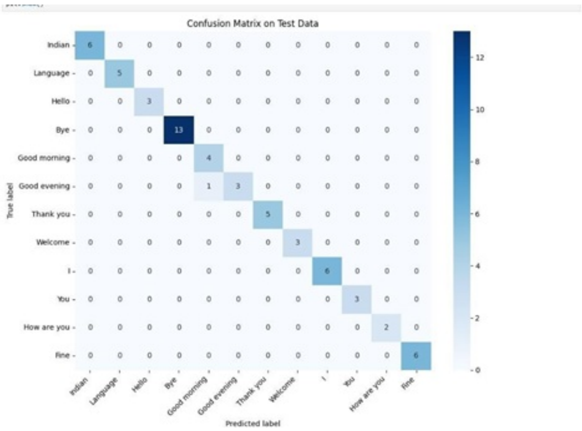


Figure 8: Confusion Matrix of 2D-CNN.

A confusion matrix (shown in Figure 8) revealed minimal misclassifications, further proving the 2D CNNs robustness. The 2D CNN also performed well in predicting gestures such as “Bye” (13/13), “Fine” (6/6), and “I” (6/6). It maintained high accuracy for “Indian” and “Language”, and showed good performance in static gestures like “You” (3/3) and “How are you” (2/2). However, it struggled slightly with “Good evening”, misclassifying one instance as “Good morning”. Moreover, the system’s strong performance can be attributed to the quality and diversity of the custom-built dataset, tailored specifically for Indian gesture styles.

The comparative analysis of the results as given in Table 1 demonstrates that the 2D CNN outperforms the 3D CNN in terms of accuracy, generalization, training efficiency, and real-time applicability. While 3D CNNs are conceptually powerful for capturing motion, their computational demands and sensitivity to data variance limit their performance on small to mid-sized datasets. The 2D CNN, however, demonstrated high reliability, minimal errors, and fast training, making it ideal for integration into the ISL to Odia text translation system. This validates our system’s current implementation and supports the use of 2D CNNs for practical sign language translation applications.

Furthermore, the successful end-to-end mapping of gestures to the Odia language not only fills a significant gap in regional assistive technologies but also represents the first known implementation of direct Indian Sign Language to Odia conversion using deep learning. The system sets a strong precedent for future work in inclusive communication tools tailored for India’s diverse linguistic landscape.

Aspect	3D CNN	2D CNN
Training Accuracy	88.62%	95.04%
Validation Accuracy	77.63%	99.56%
Test Accuracy	78.33%	98.33%
Model Complexity	High (~570K parameters)	Moderate (~160K parameters)
Training Duration	75 epochs (slow convergence)	50 epochs (fast convergence)
Strengths	Captures motion over time; useful for complex sequences	Lightweight, highly accurate, real-time ready
Weaknesses	Overfits easily, resource-heavy, and has lower generalization	Limited temporal modelling
Suitability for Real-time Deployment	Less suitable	Highly suitable

Table 1: Comparative Analysis of Performance of 3D CNN and 2D CNN Models for Gesture Recognition.

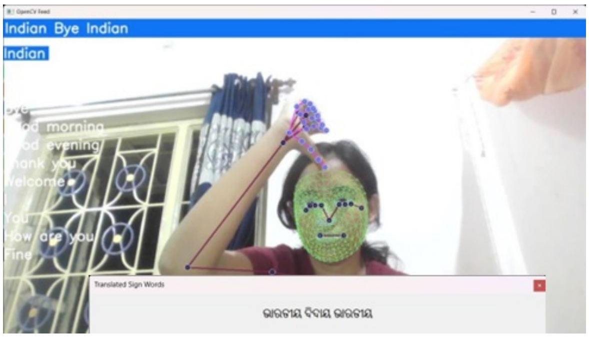
4.1 ISL to Odia Language Mapping

The model incorporates a feature that translates ISL gestures into Odia script. This is achieved through a direct mapping system using a predefined dictionary of related Odia words. Each recognized ISL gesture is mapped to its corresponding Odia word or phrase, enabling the system to provide text output in Odia.

To facilitate this, the system uses a Tkinter-based UI that allows users to seamlessly switch to Odia translation. The UI displays the translated Odia text, providing a smooth and intuitive way for users to interact with the system. This feature enhances the user experience and ensures effective communication for Odia-speaking users in the deaf and mute community. Figure 9 shows one such demo translation.

5 Conclusion and Future Scope

This work presents a comprehensive system for recognizing Indian Sign Language (ISL) gestures and translating them into the Odia language, aiming to empower the deaf and hard-of-hearing community in Odisha. By integrating computer vision and deep learning techniques, the system successfully identifies hand gestures corresponding to commonly used ISL signs and maps them to their respective Odia translations.



Indian Bye Indian

Indian

ଭାରତୀୟ ବିଦାୟ ଭାରତୀୟ

Figure 9: A Demo Translation Result.

The core of the current system utilizes MediaPipe Holistic to extract precise hand landmark coordinates, which are then processed using a Convolutional Neural Network (CNN). This combination enables efficient recognition of hand gestures captured in real time. This prototype demonstrates the potential for bridging communication gaps and improving accessibility for the hearing-impaired population, particularly those in Odia-speaking regions. Furthermore, it serves as a helpful learning tool for new individuals to learn sign language, promoting wider awareness and understanding.

In terms of model architecture, we plan to enhance temporal feature extraction using a deeper LSTM-based framework, consisting of multiple stacked LSTM layers followed by dense layers with ReLU activations for high-level feature abstraction and classification. This architecture, proven effective in prior ISL-related work, will enable our system to understand the sequence and flow of gestures more accurately, which is vital for real-time translation. We also aim to extend the system from isolated gesture recognition to sentence-level or continuous ISL translation. This will involve modelling temporal dependencies over extended gesture sequences, dynamic segmentation of signs, and restructuring of the translated output to form grammatically correct Odia sentences. This advancement will significantly improve the usability of the system in natural communication contexts.

Additionally, plans are underway to integrate the system with Odia speech synthesis, allowing the translated Odia text to be converted into voice output. This feature will make the tool more accessible, especially for users with additional literacy or visual impairments. In the long term, the system can be embedded into real-time video chat platforms to support inclusive conversations between deaf users and Odia-speaking individuals, both in-person and online.

References

Venkata Naga Sai Apurupa Amperayani, Ayan Banerjee, and Sandeep KS Gupta. 2024. Grammar-based inductive learning (gbil) for sign-spotting in continuous sign language videos. In *2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS)*. IEEE.

Sonali M Antad, Siddhartha Chakrabarty, Sneha Bhat, Somrath Bisen, and Sneha Jain. 2024. Sign language translation across multiple languages. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pages 741–746. IEEE.

Purva C. Badhe and Vaishali Kulkarni. 2015. Indian sign language translator using gesture recognition algorithm. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*. IEEE.

Aonmoy Das, Ananna Dev Aishi, Masbah Uddin Toha, and Md Fazlul Kader. 2025. Bangla sign language translator for deaf and speech impaired people using deep lstm. *International Journal of Speech Technology*, pages 1–18.

Monalisa Ghosh, Debjani De, Lovely Anand, and Satyakam Baraha. 2024. R-slr: Real-time sign language recognition system. In *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, pages 1–6. IEEE.

J Himasree, PL Jeevitha, K Deekshitha, Aashrita Kolisetty, and Soumyalatha Naveen. 2024. Video-based hand gesture recognition using random forest for sign language interpretation. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, pages 1–6. IEEE.

Tamon Kondo, Ryouta Murai, Duk Shin, and Yousun Kang. 2024. Evaluating the accuracy of real-time japanese sign language word recognition with vision transformer models trained on angular features. In *2024 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC)*, pages 1–6. IEEE.

B Natarajan, E Rajalakshmi, R Elakkiya, Ketan Kotecha, Ajith Abraham, Lubna Abdelkareim Gabralla, and V Subramaniyaswamy. 2022. Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation. *IEEE Access*, 10:104358–104374.

André Neyra-Gutiérrez and Pedro Shiguihara-Juárez. 2020. Feature extraction with video summarization of dynamic gestures for peruvian sign language recognition. In *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. IEEE.

Gadha Lekshmi P and Rohith Francis. 2024. Sign2text: Deep learning-based sign language translation system using vision transformers and phi-1.5b. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pages 282–287.

P Anantha Prabha, Mohammed Daanish, Naveen Kumar, and Nithish Kumaar. 2024. Deep-signspeak: Deep learning based sign language recognition and regional language translation. In *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–6. IEEE.

Khan Rehan and Touhid Mullick. 2023. Real time sign language translator for video conferencing platforms.

Sherin Shanavas, Naila N N, and Harikrishnan S R. 2024. Gesture recognition and sign language detection using deep learning. *International Journal of Advanced Research in Science, Communication and Technology*, 4(1):117–124.

World Federation of the Deaf. World federation of the deaf.

Kai Zhao, Kejun Zhang, Yu Zhai, Daotong Wang, and Jianbo Su. 2021. Real-time sign language recognition based on video stream. *International Journal of Systems, Control and Communications*, 12(2):158–174.

Jingchen Zou, Jianqiang Li, Xi Xu, Yuning Huang, Jing Tang, Changwei Song, Linna Zhao, Wenxiu Cheng, Chujie Zhu, and Suqin Liu. 2024. Progressive sign language video translation model for real-world complex background environments. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 519–524. IEEE.

A Appendix

- Dataset: https://drive.google.com/file/d/1fNnwoOIQP1iE68PPHQnaPgF-DgoRCdPF/view?usp=sharing
- Code: https://github.com/Astha1asn/Sign-Language-Translation-To-Odia