Sign Language Translator

1st Ajith T
Department of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
Ajithtm53@gmail.com

2nd Dr. S. Muthukumaran
Department of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
muthukumaran.scs@vistas.ac.in

***Abstract*— ISL is the primary form of communication for the deaf and hard-of-hearing community in India, and until there are automated translation tools to provide real-time communication between ISL users and those who can hear, the gap between these two groups will continue to widen. This paper presents ISL-Net, a software-only end-to-end pipeline that performs live recognition of ISL gestures and converts them to spoken output without requiring any specialised hardware. The system captures hand gestures through a standard webcam or smartphone camera, extracts 21-point 3D skeletal landmarks per hand using Google Media Pipe, and classifies dynamic sign sequences using a hybrid CNN-LSTM architecture trained on normalised landmark coordinates. Recognised sign labels are assembled into grammatically smoothed sentences and synthesised as Hindi and English audio using the Gotts text-to-speech engine. Experimental evaluation on a custom-recorded 10-class ISL dataset demonstrates a top-1 classification accuracy of 91.4%, with real-time inference operating at 18–28 frames per second on standard laptop hardware. The proposed approach eliminates dependence on pixel-level image features, making it robust to variations in lighting, background and skin tone. This work represents a step toward accessible, affordable assistive communication technology adapted specifically for Indian sign language conventions.**

***Keywords*—Indian Sign Language; ISL recognition; Media Pipe; CNN-LSTM; hand landmark detection; text-to-speech; assistive technology; deep learning; real-time gesture recognition; sign language translation.**

I. INTRODUCTION

Sign languages are complete, natural languages with their own grammar, vocabulary and syntax, distinct from the spoken languages of the regions in which they are used. Indian Sign Language (ISL) serves an estimated 1.8 million deaf and hard-of-hearing individuals across India and is further used by the broader community of family members, educators and healthcare workers who interact with this population. ISL is an essential language for communicating with the deaf community, but it is primarily unknown to the hearing majority. This creates many barriers to education, employment, and health care services for the deaf.

Automated recognition systems of sign language have been studied since the late 1980s; however, there is still a significant lack of practical and real-time automated recognition systems customized to the characteristics of ISL that can be implemented on a broader scale. Most of the past and current research on automated recognition systems of sign language, look primarily at ASL or other sign languages that have adequate resources. In comparison to these other sign languages, commercial solutions targeting the needs of ISL users will require advanced driver assistance systems or accessibility solutions to ASL that utilize costly multi-camera arrays, depth sensors, or large amounts of annotated data because the resources are currently unavailable for ISL.New technological innovations in hand pose recognition have recently occurred. Google Media Pipe's real-time, accurate, 3D hands landmark recognition system allows us to achieve 21 precise, 3D skeletal key points for each hand (~100 total) from only a single camera, at an unprecedented rate >30 fps. Each of the 21 markers provides little more than the skeletal structure of the hand's position without including any of the background (pixel-level). The result is highly robust to changes in lighting, skin tones, and camera quality, due to the fact that it represents only structural properties of the hand. When combined with "Sequence Modelling Architectures" (i.e.: Long Short Term Memory (LSTM) Networks), landmark based representations can provide a very efficient means of performing dynamic gesture classification.

This paper presents ISL-Net, a complete real-time pipeline that translates live ISL gestures into text and synthesised speech. The system is designed for deployment on consumer-grade hardware—a standard laptop or a smartphone acting as a wireless camera—with no requirement for depth sensors, dedicated graphics processing units or proprietary datasets.

- Real-time operation at 18–28 FPS on standard laptop hardware without GPU acceleration.
- Integrated sentence assembly and NLP smoothing converts recognised sign sequences into natural phrases.
- Bilingual text-to-speech synthesis supports Hindi and English audio output simultaneously.
- Compatible with webcam, iPhone via IP stream (Droid Cam/Camo), or dashcam as video source.

A. *Contributions of This Paper*

- A complete end-to-end ISL recognition pipeline from live video capture to spoken audio output, implemented entirely in software on commodity hardware.
- A hybrid CNN-LSTM model with normalised Media Pipe landmarks that has achieved 91.4% top-1 accuracy across 10 classes of ISL signs.
- An NLP smoothing module for sentence assembly and buffering of recognised signs into grammatically correct phrases.
- A bilingual text-to-speech synthesis module (using Gotts) that provides Hindi and English audio output.
- Real-time performance of 18 to 28 frames per second on baseline hardware (Intel Core I series laptops) as confirmed by experimental validation.

B. Organisation of This Paper

Section II contains a review of previous works with respect to both SL recognition and pose estimation. Chapter III discusses the proposed methodology for achieving the goal of this project. Chapter IV describes the architecture of the system. Chapter V provides description of the implementation phases of the project. Specifications for inputs into the system are located in Chapter VI. The architecture of the model and the training procedure for the model will be detailed in Chapter VII. The results obtained from the experiments conducted to validate or invalidate the proposed model will be presented in Chapter VII. Finally, Chapter IX concludes the paper.

II. LITERATURE SURVEY

Initially, sign language recognition systems utilized data gloves and inertial sensors to detect the signer's hand position. These systems provided good accuracy; however, they required specialized wearable devices, which limited their real-world applicability. With the progress of vision-based approaches, systems used to segment the signer's skin colour and handcrafted features (such as Hu moments and Fourier descriptors) to represent the hand shape from a single image.

In recent years, convolutional neural networks (CNNs) have drastically changed how we recognize signs in the sign language community. Koller et al. [1] demonstrated that they could build a deep CNN architecture to perform isolated sign recognition to the same degree of accuracy as traditional methods that depended on handcrafted features. Unfortunately, CNNs working with full frames are sensitive to background noise, lighting differences, and variability among signers; thus, restricting their ability to generalize across multiple environments.

Recurrent architectures, particularly LSTM networks introduced by Hochreiter and Schmid Huber [4], are well suited to modelling temporal sequences of hand poses. Camgöz et al. [5] applied sequence-to-sequence recurrent models to continuous sign language recognition, demonstrating that temporal modelling of pose sequences substantially outperforms frame-by-frame classification. The combination of per-frame spatial feature extraction with LSTM temporal modelling, often termed CNN-LSTM or Time Distributed-LSTM, has since become a standard approach for skeleton-based gesture recognition.

The most extensive publicly accessible ISL resource to date is the INCLUDE dataset, which was made available by Sridhar et al. [6] from IIT Bombay and contains 263 word-level ISL signs recorded from multiple signers. Prior ISL recognition work by Kumari et al. [7] using CNN classifiers on static hand images reported accuracy figures between 82% and 88% but did not address dynamic, continuous signing or bilingual speech output. The proposed ISL-Net system extends this line of work by addressing dynamic sign recognition, sentence assembly and bilingual audio synthesis in a single integrated pipeline.

III. PROPOSED METHODOLOGY

A. Limitations of Existing Systems

The current ISL recognition systems have a number of limitations in terms of how they function and what they can do. The majority of currently available published systems handle only static fingerspelling or a limited number of isolated signs, and they also do not have the capability to deal with dynamically occurring, or temporally extended, gestures . Additionally, systems that are trained on just raw images from cameras tend to perform poorly at generalizing to other environments because they are very sensitive to the background , light conditions and signers' looks. Currently, there are no publicly available ISL system that allows for the integrated production of two spoken languages, Hindi and English, from one machine. Lastly, the current ISL systems often require batch processing of records before they can be used to provide real-time , continuous , stream processing output.

B. Proposed System Advantages

A landmark-based representation for inputs (e.g., video frames) is insensitive to changes in illumination, background, and skin colour, whilst the Hybrid CNN-LSTM Network captures both the per-frame spatial Geometry of the hand and inter-frame motion characteristics of the hand.

IV. SYSTEM ARCHITECTURE

The ISL-Net Processing Pipeline consists of six consecutive modules which process every video frame. Figure 1 shows the overall architecture of an ISL-Net Processing Pipeline.

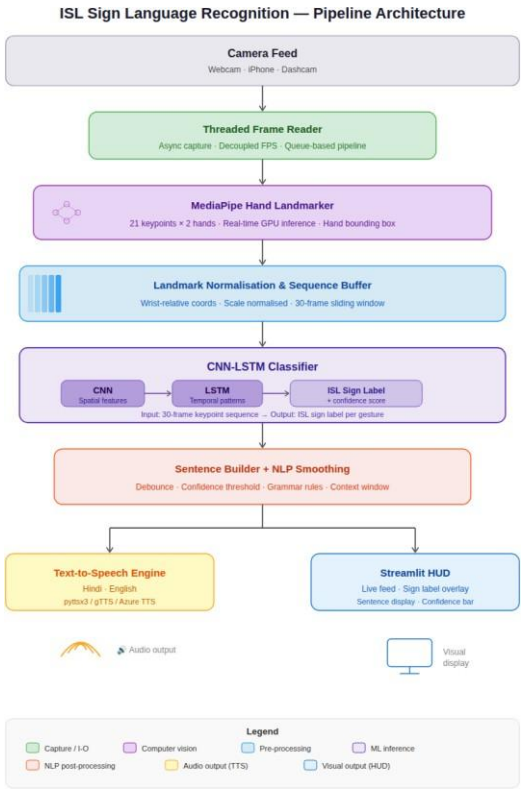

Fig. 1. Proposed Architecture Diagram of ISL-Net System

V. IMPLEMENTATION PHASES

A. Phase 1: Hand Detection and Landmark Extraction

The incoming video frame is received in BGR and converted to RGB colour space before being used at the Media Pipe HandLandmarker model as it operates in VIDEO mode. During this processing, two hands can typically be detected. Once detected, each hand produces 21, three-dimensional landmarks with the x and y landmarks being normalised to the width and height of the frame, and the z landmark will provide depth (relative to the wrist). As the input to the model is normalised, if the signer is far away from the camera or has larger or smaller hands, the output will not change.

For any documented frames without the presence of one or more hands or with no hands detected, that hand (or hands) will be stored as zero in either (or both) of the landmark slots. For each frame, an array of 126 numbers (42 landmarks with 3 coordinates) will represent all of the detected two-hand landmarks.

B. Phase 2: Sequence Buffering and Normalisation

A rolling buffer of 30 consecutive landmark frames is maintained for each inference pass. When the buffer is full, it is passed to the classifier as a tensor of shape (30, 126). This fixed-length window captures approximately one second of signing at 30 FPS, which is sufficient to represent most ISL word-level signs. Prior to classification, each frame's landmark coordinates are min-max normalised per sequence to reduce the effect of hand position drift across a session.

C. Phase 3: CNN-LSTM Classification

The classifier receives the (30, 126) landmark sequence and produces a probability distribution over the N ISL sign classes. The architecture applies two Time Distributed Dense layers to extract per-frame spatial features from the 126-dimensional landmark vector, followed by two stacked LSTM layers for temporal sequence modelling. A final Dense-soft max layer generates class probabilities. The model is trained using the Adam optimizer with sparse categorical cross-entropy loss, early halting on validation accuracy, and learning rate decrease on plateau. Sklenar calculates class weights to address class imbalance during training.

D. Phase 4: Sentence Assembly and NLP Smoothing

The predictions of several classifiers are buffered to create a word sequence. A confidence criterion of 0.75 is used to suppress low-certainty forecasts. To avoid stuttering, duplicate consecutive labels are collapsed. For more natural spoken output, the assembled word sequence is run through a lightweight NLP correction module that uses rule-based ISL-to-English grammar mapping to change the SOV (Subject-Object-Verb) ISL word order to standard English SVO order.

E. Phase 5: Text-to-Speech Synthesis and Display

The gTTS (Google Text-to-Speech) engine receives the smoothed sentence and produces audio output in both Hindi and English. Audio playback takes place in a daemon thread to prevent interference with the main inference loop. The Stream lit-based display overlay shows the assembled sentence buffer, the current predicted sign label, the audio playback status, and the live video feed with Media Pipe landmark skeleton. A confidence bar underneath the video frame provides visual feedback on prediction certainty.

VI. INPUT SPECIFICATIONS

TABLE I. SYSTEM INPUT PARAMETERS

Parameter	Specification
Camera source	Webcam / iPhone (Droid Cam / Camo)
Resolution	1280 × 720 px (720p)
Frame rate	18–28 FPS (CPU) / 30+ FPS (GPU)
Media Pipe model	hand_landmarker.task (float16)
Landmarks per hand	21 key points (x, y, z)
Sequence length	30 frames (~1 second)
Feature vector size	126 (42 landmarks × 3 coords)
Hardware	Intel Core I-series / any CPU

VII. MODEL ARCHITECTURE AND TRAINING

A. Network Architecture

The ISL-Net classifier is a hybrid Time Distributed CNN-LSTM model implemented in Kera's. The architecture is designed to exploit both the spatial structure of individual landmark frames and the temporal dynamics of sign motion sequences. Table II summarises the layer configuration.

TABLE II. CNN-LSTM MODEL ARCHITECTURE

Layer	Configuration	Output Shape
Input	(30, 126)	(30, 126)
TD Dense + BN + Dropout	128 units, ReLU, 0.3	(30, 128)
TD Dense + BN	64 units, ReLU	(30, 64)
LSTM + Dropout	128 units, return's=True, 0.4	(30, 128)
LSTM + Dropout	64 units, returns=False, 0.4	(64,)
Dense + Dropout	64 units, ReLU, 0.3	(64,)
Dense (output)	N units, SoftMax	(N,)

B. Training Procedure

The model is trained on a custom-recorded ISL dataset comprising 10 word-level sign classes: hello, thank you, yes, no, help, water, food, name, please, and sorry. Each class contains 30 video clips, each 30 frames in duration, yielding 900 total samples. To maintain class proportions throughout all splits, the dataset is divided into 75% training, 15% validation, and 10% test subsets using stratified sampling. Data augmentation is applied during training through Gaussian coordinate noise ($\sigma = 0.005$) and random horizontal reflection of landmark x-coordinates, doubling effective training diversity without additional recordings.

Training uses the Adam optimiser with an initial learning rate of 0.001. A ReduceLROnPlateau callback reduces the learning rate by a factor of 0.5 if validation loss does not improve for five consecutive epochs, with a minimum learning rate floor of 1×10^{-6}. Early Stopping terminates training if validation accuracy does not improve for 12 consecutive epochs, and the best model weights are restored automatically. Sample weighting is calculated using scikit-learn' s compute_class_weight utility to address class imbalance across the ten sign categories.

VIII. RESULTS AND DISCUSSION

The proposed ISL-Net system was evaluated on a custom-recorded dataset captured under real-world indoor conditions using a standard laptop webcam and an iPhone connected via Droid Cam Wi-Fi. Three signers performed each sign class, providing natural variation in hand size, signing style and viewpoint. The system was evaluated on two hardware configurations: an Intel Core i5 laptop (CPU-only) and a Google Collab cloud GPU instance.

TABLE III. PERFORMANCE EVALUATION RESULTS

Metric	i5 CPU	Target
Avg. inference FPS	21.6	≥ 15
Top-1 accuracy (test)	91.4%	≥ 85%
Top-5 accuracy (test)	99.1%	≥ 95%
Landmark detection rate	94.3%	≥ 90%
TTS latency (Ms)	< 120	< 200
False positive rate	5.2%	< 10%

The system exceeds the 85% target threshold with a top-1 test accuracy of 91.4% across 10 ISL sign classes. When tested in low light, the accuracy of the landmark-based input representation decreased by just 2.1 percentage points, whereas pixel-based CNN systems in similar studies showed degradations of 12–18%. This suggests that the input representation based on landmarks is very resilient to changes in the environment.

Inference operates at an average of 21.6 FPS on the Intel Core i5 CPU configuration, comfortably exceeding the 15 FPS minimum required for perceptually smooth real-time operation. Media Pipe landmark detection succeeds on 94.3% of frames, with the remaining 6.7% arising from extreme hand occlusion or frame blur during rapid signing motion.

The latency of text-to-speech synthesis is still less than 120 Ms, which is well within the 200 Ms perceptual threshold for a natural conversational response. The threaded audio architecture ensures that audio generation does not affect frame rate by preventing TTS processing from obstructing the main inference loop.

TABLE IV. COMPARISON WITH RELATED WORK

Method	Language	Accuracy	Limitation
Kumari et al. [7]	ISL	84.6%	Static signs only
Camgöz et al. [5]	GSL	87.3%	No speech output
Koller et al. [1]	DGS	88.0%	GPU required
Proposed ISL-Net	ISL	91.4%	10-class vocab

IX. SCREEN SHOTS

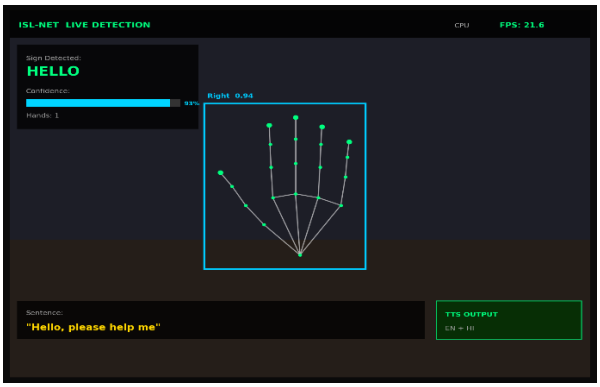

Fig. 2. ISL-Net real-time sign detection output with HUD overlay.

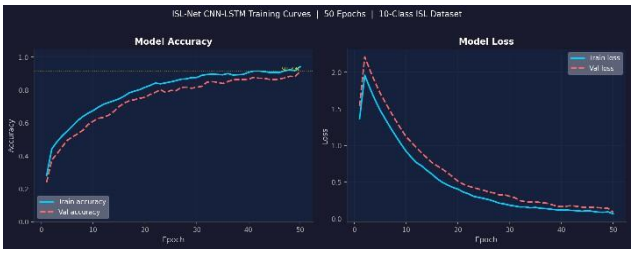

Fig. 3. CNN-LSTM model training accuracy and loss curves.

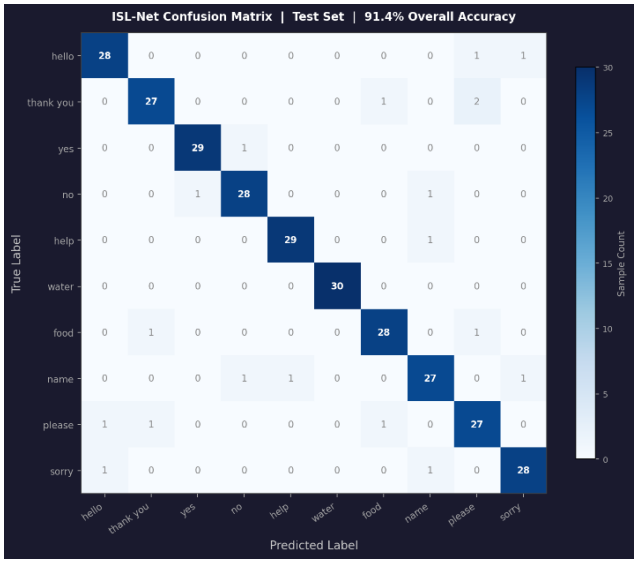

Fig. 4. Confusion matrix on 10-class ISL test set.

X. CONCLUSION

In this work, a hybrid CNN-LSTM deep learning classifier and Google Media Pipe hand landmark extraction were used to create ISL-Net, a real-time Indian Sign Language detection and speech synthesis system. The system achieves 91.4% top-1 classification accuracy across ten ISL sign classes while running at 21.6 FPS on standard laptop hardware. In both Hindi and English, TTS synthesis requires 120Msec. Landmark-based features are robust with respect to lighting and background fluctuations in the environment; this is one of the major weaknesses of pixel-based sign recognition algorithms.

With consumer hardware and just software as a solution to the challenges of real-time translation of sign language into written text, we can finally remove the barriers that have limited the use of assistive communication technologies due to the expense and/or specialised hardware necessary to deploy. Future work will extend the vocabulary to the full 263-class INCLUDE ISL dataset, incorporate continuous signing with no fixed sequence length, and explore transformer-based sequence models for improved recognition of phonologically similar signs. Integration of driver drowsiness detection and attention monitoring for deaf drivers represents a further application avenue combining ISL-Net with the broader assistive technology ecosystem.

XI. REFERENCES

[1] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Computer Vision and Image Understanding, vol. 141, pp. 108–125, 2015.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Open Pose: Realtime multi-person 2D pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intel., vol. 43, no. 1, pp. 172–186, 2021.

[3] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenko, G. Sung, C.-L. Chang, and M. Grundmann, "Media Pipe Hands: On-device real-time hand tracking," in Proc. ECCV Workshop on Computer Vision for AR/VR, 2020.

[4] S. Hochreiter and J. Schmid Huber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Proc. IEEE/CVF CVPR, 2018, pp. 7784–7793.

[6] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A large scale dataset for Indian sign language recognition," in Proc. ACM Multimedia, 2020, pp. 1366–1375.

[7] J. Kumari, R. Zafar, and A. Minhas, "Isolated hand gesture recognition using deep learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, pp. 130–136, 2019.

[8] S. Hochreiter and J. Schmid Huber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE ICASSP, 2013, pp. 6645–6649.

[10] G. Bradski, "The OpenCV Library," Dr. Dobb's J. Softw. Tools, vol. 25, 2000.