



# Real-Time Sign Language Translator Using Video Prediction

Geetha. N<sup>1</sup>, Mohanashankare.V.K<sup>2</sup>, Prathanyaa.K<sup>3</sup>

Assistant Professor, Department of Computer Applications, Vellalar College for Women, Thindal, Erode, Tamil Nadu, India<sup>1</sup>

MCA, Department of Computer Applications, Vellalar College for Women, Thindal, Erode, Tamil Nadu, India<sup>2,3</sup>

**Abstract:** Sign language is a visual language in which meaning is conveyed through hand shape, movement, body posture, facial expression, and temporal relationships between gestures. Automatic translation from video to text can reduce communication barriers between sign-language users and people who do not know the language. This paper presents a real-time sign language translator using video prediction, in which a webcam or uploaded video is processed as a sequence rather than as independent images. The proposed framework combines video acquisition, frame preprocessing, hand and pose landmark extraction, spatial feature learning, temporal sequence modelling, sign prediction, text generation, and optional speech synthesis. OpenCV can be used for frame handling, while MediaPipe provides efficient landmark extraction. CNN(Convolutional Neural Network), GRU(Gated Recurrent Unit), LSTM( Long Short-Term Memory), or Transformer models can then learn spatial and temporal patterns in the video. Recent research from 2020 to 2026 shows a transition from static and isolated-sign recognition toward continuous sign-language recognition and end-to-end sign-to-text translation. However, practical deployment remains challenging because of signer variation, occlusion, variable signing speed, limited continuous datasets, non-manual cues, temporal alignment, and computational latency. The proposed research direction therefore emphasizes efficient video prediction, robust temporal modelling, and real-time deployment, with particular relevance to Indian Sign Language.

**Keywords:** Sign Language Recognition, Video Prediction, Computer Vision, Deep Learning, MediaPipe, Temporal Modeling, Transformer, Indian Sign Language, Real-Time Translation, Speech Synthesis.

## I. INTRODUCTION

Sign language is a complete visual language used by Deaf and hard-of-hearing communities. Unlike spoken-language recognition, sign-language understanding depends on coordinated visual cues, including hand configuration, orientation, location, movement, body posture, facial expression, and head movement. A computer system that recognizes only a single image may identify a hand shape, but it cannot fully model the temporal evolution of a dynamic sign. Consequently, video-based sign-language recognition is a sequence-learning problem in which spatial and temporal information must be considered together.

Research during 2020–2026 demonstrates this progression. Mistree et al. [1] showed that frame-sequence generation and augmentation can improve Indian Sign Language video recognition. Abdul et al. [2] combined attention-based Inception features with BiLSTM temporal modelling for real-time sign classification. Subramanian et al. [3] introduced a MediaPipe-optimized GRU model for Indian Sign Language, illustrating the usefulness of landmark representations for efficient sequential learning. Musthafa and Raji [4] investigated a real-time Indian Sign Language recognition prototype. Transformer-based research subsequently strengthened spatiotemporal modelling. A full Transformer network for word-level recognition [5] integrated visual and temporal self-attention, while recent continuous recognition studies use multi-scale temporal perception [6] and adaptive video representations [7].

The objective of the proposed research is to design an interactive video-to-text and text-to-speech system that can operate from a standard camera. The system is intended to capture a sign sequence, extract meaningful visual features, predict the sign or sign sequence, and display the recognized output with a confidence value. The approach is particularly relevant to web-based accessibility applications because the same architecture can support both uploaded videos and live webcam streams.

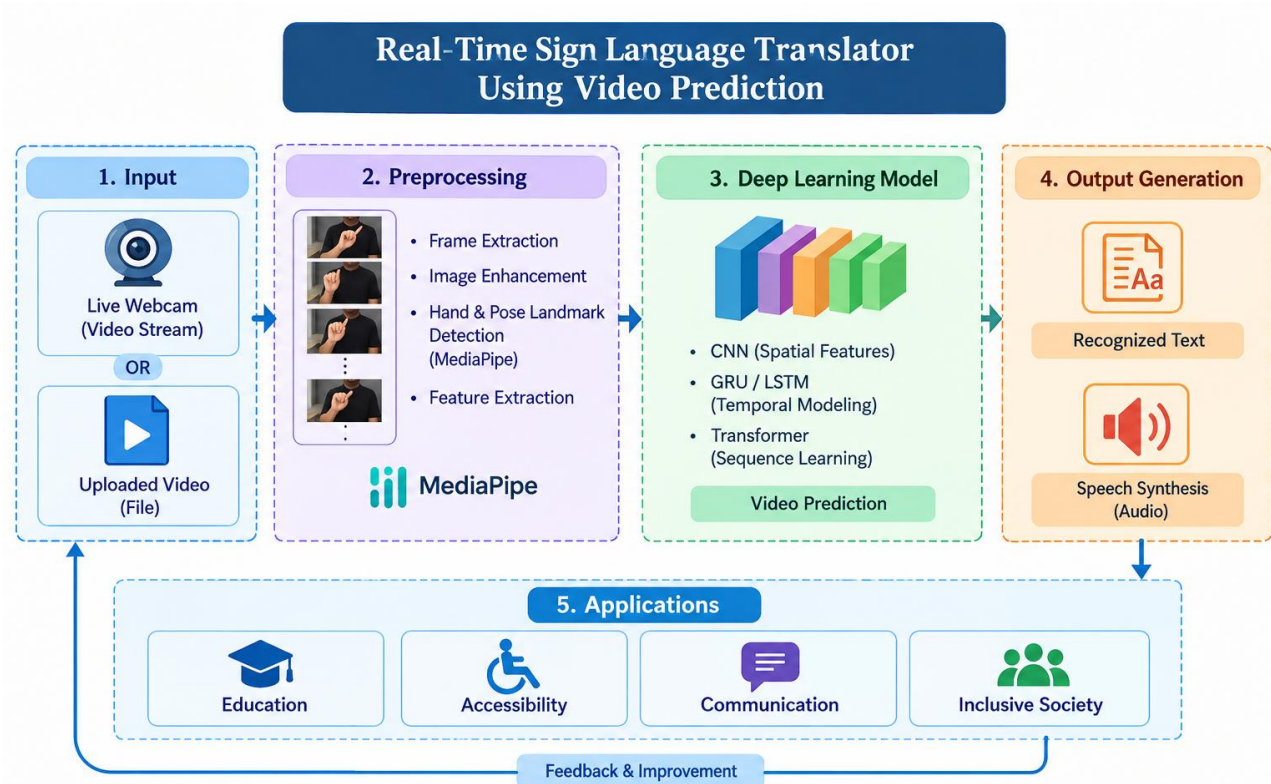


Fig. 1. Overall system architecture of the proposed real-time sign language translator.

## II. LITERATURE SURVEY

The literature can be organized into four related directions: isolated or static sign classification, landmark-based temporal recognition, Transformer-based video recognition, and continuous sign-language translation. The first direction establishes strong spatial feature extraction, but often assumes controlled data and a fixed sign vocabulary. The second direction represents gestures using hand or pose landmarks and applies sequence models such as GRU or LSTM, reducing the computational burden of processing every raw pixel. The third direction uses self-attention to model longer temporal dependencies and capture relationships among visual regions. The fourth direction attempts to map continuous sign videos directly into natural-language sentences.

Mistree et al. [1] used Indian Sign Language video as input and applied frame-sequence generation and image augmentation. Their work is important because it shows that temporal sampling and data augmentation can improve performance when sign-language datasets are limited. Abdul et al. [2] addressed real-time recognition by combining an attention-based Inception CNN with BiLSTM. Their model considered variable sign duration, dynamic backgrounds, lighting, and camera distance, highlighting the practical requirements of real-time systems.

Subramanian et al. [3] integrated MediaPipe with an optimized GRU model. The study specifically discusses hand tracking, occlusion, computational cost, and sequential information processing. This is closely related to the proposed system because MediaPipe-style landmarks can be extracted quickly and supplied to a temporal network. Musthafa and Raji [4] demonstrated a real-time Indian Sign Language recognition system, reinforcing the feasibility of converting visual signs into machine-readable outputs in an interactive setting.

El-Alfy and Luqman [8] surveyed acquisition devices, databases, recognition techniques, fingerspelling, isolated signs, and continuous sentence recognition. Their taxonomy makes clear that the field involves more than classification: acquisition, segmentation, recognition, translation, and linguistic structure must be addressed together. Alyami et al. [9] later reviewed 25 years of continuous sign-language recognition and identified issues involving data diversity, modalities, non-manual features, temporal modelling, and alignment. Shahin and Ismail [10] examined the transition from rule-based approaches to deep-learning Transformer architectures for sign-language translation and emphasized the requirements of real-time quality-of-service systems.



More recent research focuses on efficient and generalizable video models. Swin-MSTP [6] uses a Swin Transformer for spatial representation and multi-scale temporal perception for continuous sign recognition. Liu et al. [7] investigated adaptive video representation enhancement for end-to-end sign-language translation, aiming to convert continuous videos into coherent natural-language sentences without an intermediate gloss sequence. Novopoltsev et al. [11] explored pre-training for video-based sign-language learning, indicating a movement toward foundation-model approaches. Toshpulatov et al. [12] surveyed deep-learning pathways for recognition, translation, and production. Finally, Al-Mohair and Al-Qershi [13] reviewed video-based SLR in the deep-learning era and highlighted representation, temporal modelling, alignment, benchmarking, reproducibility, and deployment as interconnected issues. A 2026 review by Alshammari et al. [14] further synthesizes video-based deep-learning research from 2020 to 2026, including spatiotemporal extraction, temporal modelling, attention, motion representations, transfer learning, and hybrid frameworks.

TABLE SUMMARY OF RECENT SIGN LANGUAGE VIDEO RESEARCH

| Author & Year                    | Observation                                        | Inference                                  |
|----------------------------------|----------------------------------------------------|--------------------------------------------|
| Mistree et al. [1], 2020         | ISL video; frame sequences and augmentation        | Video preprocessing and dataset expansion  |
| Abdul et al. [2], 2021           | Attention Inception + BiLSTM                       | Real-time spatial-temporal modelling       |
| Subramanian et al. [3], 2022     | MediaPipe + optimized GRU                          | Efficient landmark-based sequence learning |
| Musthafa & Raji [4], 2022        | Real-time ISL recognition                          | Interactive recognition pipeline           |
| Full Transformer [5], 2022       | Vision Transformer + temporal Transformer          | Self-attention for video sequences         |
| El-Alfy & Luqman [8], 2022       | Survey and taxonomy                                | Datasets, acquisition and recognition      |
| Alyami et al. [9], 2024          | 25-year CSLR survey                                | Continuous recognition gaps                |
| Liu et al. [7], 2024             | Adaptive video representation + Transformer        | End-to-end sign translation                |
| Swin-MSTP [6], 2025              | Swin Transformer + multi-scale temporal perception | Variable-duration temporal modelling       |
| Al-Mohair & Al-Qershi [13], 2026 | Video SLR review                                   | Deployment and reproducibility challenges  |

### III. RESEARCH GAP

Although recent systems report strong results on benchmark datasets, several gaps remain before a sign-language translator can operate reliably in everyday environments. First, many datasets are limited in signer diversity, vocabulary, camera viewpoints, or environmental conditions. A model trained on a small and homogeneous dataset may not generalize to new users. Second, continuous sign recognition requires temporal segmentation and alignment because the system must determine when one sign ends and another begins. Third, hand-only models can miss facial expressions, mouth movements, head orientation, and body posture that contribute to meaning. Fourth, large Transformer architectures may require substantial memory and processing resources, creating a trade-off between recognition quality and real-time latency.

A further gap exists between recognition and translation. Recognizing a word or gloss is not equivalent to generating a grammatically correct natural-language sentence. End-to-end translation research [7] addresses this problem, but it generally requires large datasets and substantial computation. For Indian Sign Language, there is also a need for larger continuous datasets representing different signing styles and real-world environments. These limitations motivate a lightweight architecture that can combine robust visual or landmark features with temporal attention and language-level post-processing.

### IV. MOTIVATION

The main motivation is to develop an accessible communication tool using hardware that is already available to many users. A webcam-based system does not require special gloves or wearable sensors and can therefore be implemented as a desktop or web application. The proposed pipeline can combine OpenCV for video capture and frame



management, MediaPipe for hand landmark detection, and a trained temporal model for prediction. The result can be displayed as text and optionally converted into speech using a text-to-speech module.

The research is also motivated by the need to evaluate real-time systems using more than recognition accuracy. Useful measures include word or sign error rate, precision and recall, signer-independent performance, inference time, frames per second, memory usage, and robustness under changes in lighting, background, occlusion, and signing speed. Such evaluation connects a research prototype with the requirements of an actual assistive application.

## V. PROPOSED SYSTEM

### A. Video Acquisition and Preprocessing

The system receives a live webcam stream or an uploaded sign-language video. Frames are sampled at a controlled rate and resized or normalized. Optional background reduction and region-of-interest extraction can reduce irrelevant visual information.

### B. Landmark and Spatial Feature Extraction

MediaPipe Hands or a comparable pose-estimation framework can identify hand landmarks. Optional face and body landmarks can preserve non-manual cues. Alternatively, an RGB feature extractor such as CNN, Vision Transformer, or Swin Transformer can learn appearance features directly from frames.

### C. Temporal Video Prediction

The extracted feature sequence is supplied to a GRU, LSTM, temporal convolutional network, or Transformer. The temporal model learns movement patterns and relationships across consecutive frames. Attention can emphasize frames that contain the most informative hand or facial movements.

### D. Text and Speech Generation

The prediction layer produces a sign label, gloss sequence, or sentence. A language-processing stage can improve sentence readability, and a text-to-speech module can generate audio. The user interface can display recognized text, confidence score, processing time, and audio controls.

## VI. APPLICATIONS

The proposed technology can be applied in communication assistance, sign-language education, public-service kiosks, customer-support systems, accessibility features for websites, smart-home interfaces, and human-computer interaction. Educational applications can provide immediate feedback while learners practice signs. Public-service applications can assist communication at reception desks or information counters. In web applications, a webcam-based interface can allow users to communicate without specialized hardware. However, automatic translation should be considered assistive technology in high-stakes situations, where qualified human interpretation may still be necessary.

## VII. CONCLUSION

Real-time sign language translation using video prediction is an interdisciplinary problem combining computer vision, deep learning, temporal sequence modelling, and natural-language processing. The literature from 2020 to 2026 demonstrates a clear progression from static and isolated-sign recognition toward landmark-based temporal models, Transformer architectures, continuous recognition, and end-to-end sign-to-text translation. The proposed framework builds on these developments by integrating video capture, preprocessing, landmark or visual feature extraction, temporal prediction, text generation, and speech synthesis.

The main contribution of the proposed research direction is its focus on a practical real-time pipeline rather than recognition accuracy alone. The system should be designed for low latency, signer-independent evaluation, robustness to environmental variation, and uncertainty-aware prediction. A diverse dataset and careful temporal modelling are essential because a sign is defined not only by its appearance but also by its movement and context. With these considerations, the proposed system can provide a strong foundation for an accessible video-based sign-language translation application.



## VIII. FUTURE DIRECTIONS

Future research can investigate larger continuous Indian Sign Language datasets containing multiple signers, regional variations, different backgrounds, and varied camera viewpoints. Multimodal learning can combine hand, body, face, and motion information. Self-supervised and transfer-learning methods may reduce the dependence on expensive frame-level annotation. Lightweight Transformer, GRU, or hybrid architectures can be explored for browser and edge-device deployment. Further work can address continuous sign-to-text translation, multilingual output, personalization for individual signing styles, calibrated confidence estimation, and privacy-preserving on-device inference.

Another important direction is evaluation with standardized metrics and realistic test conditions. Future studies should report recognition or translation quality together with inference latency, memory use, frames per second, and cross-signer generalization. Collaboration with sign-language users and domain experts can also improve dataset design, interface usability, and the linguistic validity of generated translations.

## REFERENCES

- [1]. K. Mistree, D. Thakor, and B. Bhatt, "Towards Performance Improvement in Indian Sign Language Recognition," in Proceedings of the 17th International Conference on Natural Language Processing (ICON), pp. 349–354, 2020.
- [2]. W. Abdul, M. Alsulaiman, S. U. Amin, M. Faisal, G. Muhammad, F. R. Albogamy, M. A. Bencherif, and H. Ghaleb, "Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM," Computers & Electrical Engineering, vol. 95, art. 107395, 2021, doi: 10.1016/j.compeleceng.2021.107395.
- [3]. B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An integrated MediaPipe-optimized GRU model for Indian sign language recognition," Scientific Reports, vol. 12, art. 11964, 2022, doi: 10.1038/s41598-022-15998-7.
- [4]. N. Musthafa and C. G. Raji, "Real time Indian sign language recognition system," Materials Today: Proceedings, vol. 58, pp. 504–508, 2022, doi: 10.1016/j.matpr.2022.03.011.
- [5]. "Full transformer network with masking future for word-level sign language recognition," Neurocomputing, vol. 500, pp. 115–123, 2022, doi: 10.1016/j.neucom.2022.05.051.
- [6]. S. Alyami, H. Luqman, and M. Hammoudeh, "Swin-MSTP: Swin transformer with multi-scale temporal perception for continuous sign language recognition," Neurocomputing, vol. 617, art. 129015, 2025, doi: 10.1016/j.neucom.2024.129015.
- [7]. Z. Liu, J. Wu, Z. Shen, X. Chen, Q. Wu, Z. Gui, L. Senhadji, and H. Shu, "Improving End-to-End Sign Language Translation With Adaptive Video Representation Enhanced Transformer," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 9, pp. 8327–8342, 2024, doi: 10.1109/TCSVT.2024.3376404.
- [8]. E. S. M. El-Alfy and H. Luqman, "A comprehensive survey and taxonomy of sign language research," Engineering Applications of Artificial Intelligence, vol. 114, art. 105198, 2022, doi: 10.1016/j.engappai.2022.105198.
- [9]. S. N. Alyami, H. Luqman, and M. Hammoudeh, "Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects," Information Processing & Management, vol. 61, no. 5, art. 103774, 2024, doi: 10.1016/j.ipm.2024.103774.
- [10]. N. Shahin and L. Ismail, "From rule-based models to deep learning transformers architectures for natural language processing and sign language translation systems: Survey, taxonomy and performance evaluation," Artificial Intelligence Review, vol. 57, art. 271, 2024, doi: 10.1007/s10462-024-10895-z.
- [11]. M. Novopoltsev, A. Tulenkov, R. Murtazin, R. Akhidov, I. Zemtsova, E. Bojarskaja, D. Bondarenko, A. V. Savchenko, and I. Makarov, "Video-based learning of sign languages: One pre-train to fit them all," in IEEE International Conference on Data Mining Workshops (ICDMW), pp. 899–902, 2024, doi: 10.1109/ICDMW65004.2024.00124.
- [12]. M. Toshpulatov, W. Lee, J. Jun, and S. Lee, "Deep learning pathways for automatic sign language processing," Pattern Recognition, vol. 164, art. 111475, 2025, doi: 10.1016/j.patcog.2025.111475.
- [13]. H. K. Al-Mohair and O. Al-Qershi, "Video-based sign language recognition in the deep learning era: Representations, temporal modelling, alignment, benchmarks, and open challenges," The Visual Computer, vol. 42, art. 457, 2026, doi: 10.1007/s00371-026-04665-x.
- [14]. Alshammari et al., "A comprehensive review of deep learning approaches for video-based sign language recognition: Datasets, challenges and insights," Multimodal Technologies and Interaction, vol. 10, no. 6, art. 58, 2026, doi: 10.3390/mti10060058.