International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 10 | Oct - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Smart Sign Language Interpreter for Text and Speech Conversion

Dr.K.Malarvizhi,Assistant Professor,B.Tech (IT) & CIT

R.Vijaykumar,Student,B.Tech(IT) & CIT

B K.Pavishnu,Student,B.Tech(IT) & CIT

P.Hariharan,Student,B.Tech(IT) & CIT

A.Kavin,Student,B.Tech(IT) & CIT

Abstract - The communication obstacle for individuals with hearing impairments causes severe issues in education and workplaces. This research considers how artificial intelligence (AI), such as computer vision and language models, can be utilized to develop an intelligent sign language interpretation system. This system is able to rapidly translate gestures into text and speech. We surveyed recent progress from 2018 to 2025 in gesture recognition, deep learning, and natural language processing. We have structured the work into four broad categories: sensor-based motion capture, AI-based gesture recognition, text-to-speech translation frameworks, and real-time communication interfaces. Our review emphasizes the strengths and limitations of today's systems. Accuracy can go up to 95% in laboratory conditions, but real-time applications in ever-changing environments are still not possible. This project envisages a straightforward design that integrates vision-based gesture recognition, predictive AI interpretation, and speech generation to bridge the communication divide smoothly. The research also identifies areas where additional research is required, including support for multiple languages, context-sensitive interpretation, and low-latency real-time execution. This paper hopes to develop an inclusive and smart communication aid for the hearing-impaired community. Keywords: AI, Sign Language Recognition, Real-Time Translation, Gesture Detection, Speech Synthesis, Hearing-Impaired Communication, Accessibility Technology.

***Key Words*: AI, hand gestures, classification, instant translation, sign language recognition, speech production, connection with the deaf community, assistive technology.**

1.Introduction

Human interaction is all about human communication. Yet, millions of deaf people live daily hardships in school, in the office, and during social interaction. Human sign language interpreters, the conventional communicators, are most likely not available, out of budget, and not present in real-time. There is a need for automatic systems to fill this communication gap and be able to efficiently and accurately substitute.

The recent development in artificial intelligence (AI), computer vision, and language models has enabled us to create systems capable of identifying sign language gestures and translating them into text or speech in real time. Depending on what they see through hand movement, facial movement, and body movement, these AI systems read and interpret sign language. It makes it more convenient to facilitate better communication among deaf people.

Notwithstanding the achievements, existing systems still exhibit decreased accuracy under dynamic conditions, constrained gesture vocabularies, and latency in real-time translation. In this article, an artificial intelligence-based sign language interpreter integrating gesture recognition, predictive translation, and speech synthesis is suggested. The app seeks to ensure universal, consistent, and effective communication, enhancing accessibility and enabling interaction among deaf and hard-of-hearing people and society in general.

2. Technologies Foundations

2.1 Machine Learning

Machine Learning (ML) is a technique that is used to identify gestures. The method involves learning the patterns from the data collected by the camera of a smartphone, which is labeled. Support Vector Machines (SVMs),K-Nearest Neighbors(KNN), and Random Forests are typical examples of algorithms that can be utilized for static gestures. Feature extraction or breaking down the features of the hand, the positions, angles, and the movements in pictures has been captured by the camera of the smartphone is the concept behind the use of the feature extraction method. Some of the preprocessing techniques, like normalization and dimensionality reduction, can significantly increase the accuracy of a model. ML is a good candidate for smaller datasets or can be a suitable low-computation environment

such as a smartphone. Feature engineering also contributes to better results. The ML models are designed in such a way that they can be trained quickly and can be executed efficiently on mobile devices thus being a good option for vocabularies of limited gestures. In fact, ML lays the foundation for the recognition of gestures on mobile phones and is characterized by a very short time between the recognition and the mobile use.

2.2 Deep Learning

Deep Learning (DL) is a tool that is capable of automatically extracting its features from images or videos captured by a smartphone camera without any human intervention. Convolutional Neural Networks(CNNs) find out the shapes and positions of hands. For capturing the temporal dependencies, RNNs, LSTMs, or GRUs are used for sequential modeling of gestures. Attention-based Transformer models help easy decoding of gesture sequences that are context-aware. Using transfer learning with pre-trained models like MediaPipe Hands or OpenPose not only makes the training time shorter but also the model more accurate. The procedure of data augmentation, for example, rotation, flipping, and scaling, not only makes the model strong but also becomes compatible with lighting and orientation changes. DL is simply terrific at continuous gesture recognition on smartphones. Also, the synergy with NLP allows the recognized gestures to be converted into text and sentences. Even though DL is a resource-hungry technology, it can still be done on contemporary smartphones with GPUs or NPUs. In fact, it is the one that provides high accuracy and real-time performance for mobile deployment.

2.3 Natural Language Processing(NLP)

NLP changes the recognized gestures into readable and grammatically correct text. Sequence-to-sequence models and large language models (LLMs) such as GPT or BERT provide the translations that are the most accurate. The use of context-aware translation, in fact, leads to the retention of the original meaning of multi-word phrases. Multilingual NLP models make it possible for direct translation into different spoken languages on mobile devices. The support of gesture recognition is the integration that allows the end-to-end communication. Text normalization is one of the features that contribute to the readability of the text. Real-time processing on mobile devices is the one that provides very low latency. NLP enables the formation of new sentences using continuous gestures. Mobile friendly NLP frameworks are the ones that are responsible for the adjustment of the computational load. In general, NLP is indeed a very effective method of linking the gestures to the human language for mobile users.

2.4 Text-to-Speech (TTS) and Accessibility Tools

Deep Learning (DL) is a tool that is capable of automatically extracting its features from images or videos captured by a smartphone camera without any human intervention. Convolutional Neural Networks (CNNs) find out the shapes and positions of hands. For capturing the temporal dependencies, RNNs, LSTMs, or GRUs are used for sequential modeling of gestures. Attention-based Transformer models help easy decoding of gesture sequences that are context-aware. Using transfer learning with pre-trained models like MediaPipe Hands or OpenPose not only makes the training time shorter but also the model more accurate. The procedure of data augmentation, for example, rotation, flipping, and scaling, not only makes the model strong but also becomes compatible with lighting and orientation changes. DL is simply terrific at continuous gesture recognition on smartphones. Also, the synergy with NLP allows the recognized gestures to be converted into text and sentences. Even though DL is a resource-hungry technology, it can still be done on contemporary smartphones with GPUs or NPUs. In fact, it is the one that provides high accuracy and real-time performance for mobile deployment.

2.5 Smartphone Cameras and Sensors

The smartphone's built-in camera captures hand gestures in real time. High-resolution cameras enhance detection of hand joints, finger positions, and orientations. Dual-lens cameras or computational methods can approximate depth estimation. The camera input is processed frame by frame to recognize gestures. Image preprocessing, including resizing, normalization, and background filtering, boosts accuracy. Smartphone sensors, like accelerometers, can optionally improve motion tracking. Mobile-friendly computer vision frameworks, such as MediaPipe, efficiently handle hand keypoints. Real-time detection ensures low-latency translation. The system is optimized to function with limited battery and processing power. Cameras serve as the main data acquisition tool in this smartphone-based setup.

2.6 Deep Learning Frameworks

Mobile-compatible DL frameworks, like TensorFlow Lite or PyTorch Mobile, are used for model deployment. CNNs and RNNs are trained and optimized for on-device inference. Transfer learning permits the use of pre-trained models for gesture recognition. DL frameworks support real-time processing of camera input on smartphones. Model quantization reduces size and computation for efficient mobile performance. Integration with smartphone OS APIs ensures seamless camera and TTS access. Frameworks support multi-threading and GPU acceleration on smartphones. Preprocessing pipelines are optimized for mobile hardware. Continuous gesture sequences are processed efficiently in real time. Frameworks enable full end-to-end gesture-to-speech translation on mobile devices.

3. Machine Learning Algorithms and Deep Learning Algorithms

3.1 Support Vector Machines (SVM)

Support Vector Machines classify static hand gestures captured by smartphone cameras. They find the best hyperplane to separate gesture classes in feature space. Hand positions, angles, and other geometric features are extracted and fed to the model. SVM is effective for small datasets and simple tasks. It offers fast training and prediction, making it suitable for on-device smartphone use. Its high accuracy for limited gesture vocabularies makes it a dependable baseline model.

3.2K-Nearest Neighbors(KNN)

KNN classifies gestures based on similarity to nearby examples in the feature space.It uses distance metrics like Euclidean distance to identify the closest gesture samples. KNN does not require an explicit training phase and is easy to implement on smartphones. It performs well with small datasets and can manage multi-class gesture recognition. Preprocessing and feature normalization improve its performance. KNN validates predictions and serves as a simple, interpretable model.

3.3 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees for improved classification accuracy. Each tree is trained on a subset of gesture features, and majority voting determines the final prediction. It is resistant to overfitting and can handle noisy data captured by smartphone cameras. Random Forest suits static gesture recognition with limited computational resources. Analyzing feature importance can optimize the system. It offers reliable results for low-complexity gestures in real-time mobile settings.

3.4 Convolutional Neural Networks(CNNs)

CNNs automatically extract spatial features from hand images captured by smartphone cameras. Convolution and pooling layers detect edges, shapes, and finger positions. CNNs are highly effective for static gesture recognition and image classification.Using transfer learning with pre-trained CNN models like MobileNet reduces training time.CNNs are optimized for real-time inference on smartphones through frameworks like TensorFlow Lite. They provide high accuracy even in varied lighting and backgrounds.

3.5Recurrent Neural Networks(RNNs) and LSTM

RNNs and LSTMs are utilized for recognizing sequential gestures in continuous sign language.hey capture temporal dependencies between consecutive gestures, maintaining context over time. LSTMs manage long-term dependencies to avoid information loss in gesture sequences. These models allow for smooth translation of multiple gestures into coherent text. They can be optimized for smartphone deployment with lightweight architectures. RNNs/LSTMs ensure accurate interpretation of dynamic signs in real-time.

3.6 Transformers and Attention Mechanisms

Transformers with attention mechanisms model the relationship between gestures in a sequence, allowing context-aware translation. They process gesture sequences in parallel and focus on important movements to create accurate text. Attention mechanisms enable the system to prioritize critical gestures during translation. Transformers are especially useful for complex sentences or phrases in sign language. Optimized transformer models can operate on modern smartphones using mobile DL frameworks

4.Comparative analysis

Paper / System	Technology Used	Algorithm / Model	Features	Accuracy / Results
Madhiarasan, 2022	Vision-based SLR (survey of methods)	Survey of CNNs, RNNs, Transformers, sensor-based methods	Broad overview of modalities, datasets, open challenges and future directions — good for literature grounding.	Summarizes state-of-the-art ranges (isolated SLR often 80–95% on constrained datasets).
WLASL — Word-level Large-scale ASL Dataset (Li et al., 2020 / WACV)	Large video dataset for ASL	Baseline: appearance-based CNNs, pose-based Pose-TGCN (temporal GCN)	Largest public word-level ASL dataset (2000+ signs), benchmark for word-level recognition.	Baseline top-10 accuracies reported; large-scale challenges remain (pose & appearance models gave comparable performance).
RWTH-PHOENIX-Weather 2014T (Forster et al.)	Video corpus for German Sign Language (continuous)	Seq2seq models for sign→text; CTC-based pipelines	Widely used benchmark for continuous sign-language translation (video→text) and model evaluation.	Enabled early SLR→SLT systems; reported baselines and established dataset for translation tasks.
DeepASL (Fang, Co, Zhang, 2018)	Infrared sensing + vision hybrid	Hierarchical bidirectional RNN (HB-RNN) + CTC for sentence-level translation	Non-intrusive sensing design; supports word & sentence level translation; addresses lighting/occlusion issues.	Reported ~94.5% word-level accuracy on 56 words; ~8.2% WER on unseen sentences
Núñez-Marcos et al., 2022/2023	Survey: SLR → SLT pipelines	Reviews classical and neural translation architectures	Deep review specifically of machine translation aspects (glossing, alignment, BLEU evaluation issues).	Summarizes translation metrics (BLEU/ROUGE) and gaps in continuous translation.
Liang et al., 2023 — MDPI	Survey of SLT/SLR methods	Covers CNNs, Transformers, pose-based and multimodal solutions	Useful modern summary linking recognition → translation → avatar rendering & evaluation.	Presents comparative tables of methods and dataset results across tasks.
Kumar et al., 2024	Vision + LLM integration (paper prototype)	Sign recognition (classifiers) + LLM-based translation/refinement	Proposes bridging ASL and ISL via LLM-based grammatical/context refinement; an explicit LLM-in-the-loop pipeline.	Reports improved translation fluency (authors claim BLEU/fluency gains over baselines).
Srivastava et al., 2024	MediaPipe Holistic (landmarks) + deep learning	Pose/landmark features + sequence models (CNN/LSTM/Transformer variants)	Practical pipeline for ISL continuous recognition using lightweight landmark features (good for mobile/prototyping).	Authors report competitive results for ISL on their dataset (suitable for real-time prototypes).
system / H2020 & publications	Multi-camera vision + structured pipelines	Pose estimation + ML mapping → text	One of the first deployed systems for (demo) sign→text in public spaces; industrial prototype.	Demonstrated high accuracy in controlled demos; requires special hardware setup.

5. Challenges and Future Directions

5.1 Current challenges

5.1.1 Limited Datasets

Most sign language datasets are small and specific to particular languages. Regional languages like ISL often lack publicly available datasets. The collection and annotation of data require significant time and effort. Small datasets decrease the AI's accuracy and make it hard to generalize across different users.

5.1.2 Gesture Variability

Different people perform the same sign in varied ways, including speed, style, and hand shape. Facial expressions and body posture can alter meanings. Recognizing these differences is tough for AI models, leading to possible misinterpretations in real situations. Ensuring consistent recognition is a challenge.

5.1.3 Continuous vs. Isolated Signs

Recognizing isolated words is easier than interpreting full sentences. Continuous signing needs precise segmentation of gestures. Pauses, overlaps, and transitions complicate sentence-level translation. AI often struggles to maintain context during conversations.

5.2 Future Directions

5.2.1 Multi-Language Support

Future systems may include multiple sign languages, making communication accessible across the globe. Expanding datasets is crucial for training purposes and can benefit users from diverse backgrounds. This approach promotes universal inclusivity in communication.

5.2.2 Context-Aware Interpretation

AI can learn to capture facial expressions and body cues. Understanding context improves translation accuracy, allowing sentence-level meaning to be preserved. This enhances the natural flow of communication while reducing errors from misinterpretation.

5.2.3 Lightweight, Interactive Systems

Optimized models may be executed on mobile or low-power devices to support smoother near real-time translations. This facilitates interactive conversation between hearing and hearing-impaired individuals. This improves usability in real-life environments and leads to adoption in daily life.

6. Conclusion

AI sign language translation can minimize communication differences between the hearing impaired and others by a significant margin. This project can demonstrate that it is possible to have a system where gesture recognition and language processing come together in an effort to translate sign language to speech and text and offer effective communication. Despite there being constraints in the form of poor data, gesture variation, and correct sentence translation, adaptive AI techniques can be employed to leverage against these. Enhanced augmentation for support of multiple languages, contextual translation, and light-weight interactive systems will also enhance the ease of use and usage level. In general, sign language interpreters supported by AI are a step in the right direction towards simpler and more accessible communication for the deaf and hard-of-hearing.

References

S. Madhiarasan, "A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets," arXiv preprint arXiv:2204.03328, 2022. [Online]. Available: https://arxiv.org/abs/2204.03328

X. Li, N. Thakur, C. Thorne et al., "Word-Level American Sign Language (WLASL) Dataset and Baseline Models," Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2020. [Online]. Available: https://github.com/dxli94/WLASL

J. Forster, C. Schmidt, O. Koller et al., "RWTH-PHOENIX-Weather 2014T: Parallel Corpus for Continuous Sign Language Translation," Proc. 9th International Conference on Language Resources and Evaluation (LREC), 2014. [Online]. Available: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-weather-2014T/

Z. Fang, C. Co, and Q. Zhang, "DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation," Proc. ACM Int. Conf. on Mobile Systems, Applications, and Services (MobiSys), 2018. [Online]. Available: https://arxiv.org/abs/1802.07584

A. Núñez-Marcos, J. Domínguez, and L. Suárez, "A Survey on Sign Language Machine Translation," Expert Systems with Applications, vol. 207, 2022. [Online]. Available: https://doi.org/10.1016/j.eswa.2022.118993

Z. Liang, H. Wang, and Y. Zhao, "Sign Language Translation: A Survey of Approaches and Methods," Electronics, vol. 12, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2678

M. Kumar, S. S. Visagan, T. Mahajan and A. Natarajan, "Enhanced Sign Language Translation between American Sign Language and Indian Sign Language Using LLMs," arXiv preprint arXiv:2411.12685, 2024. [Online]. Available: https://arxiv.org/abs/2411.12685

A. Srivastava, A. Singh, and P. Verma, "Continuous Sign Language Recognition System using Deep Learning with MediaPipe Holistic," arXiv preprint arXiv:2411.04517, 2024. [Online]. Available: https://arxiv.org/abs/2411.04517

SignAll Technologies Inc., "SignAll: Automatic Sign Language Translation System," 2019. [Online]. Available: https://signall.us

"Silence Speaks: Text-to-Sign Avatars in Real-World Trials," Wired Magazine, 2025. [Online]. Available: https://www.wired.com/story/sign-language-avatar-silence-speak