IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154253

Sign Language Recognition Using Deep Learning-A Review

Karthik Reddy A[1], Kushal Reddy K[2], Mallikarjuna S[3], P Akash Patil[4],

Dr. Muhibur Rahaman T.R[5]

6th Sem B.E.(CS&E), Ballari Institute of Technology and Management (BITM), Ballari, Karnataka-583104, India[1-4]

Associate Professor, Department of Computer Science and Engineering,

Ballari Institute of Technology and Management (BITM), Ballari, Karnataka 583104, India[5]

Abstract: For millions of people who are deaf or hard of hearing, sign language is more than just a communication tool it is the very foundation of their identity and daily life. Yet, most of the hearing world cannot understand it, which creates a serious and ongoing barrier to education, employment, and basic social participation. This paper looks at how deep learning can help close that gap through automated Sign Language Recognition (SLR). We reviewed systems that range from simple rule-based approaches all the way to the latest transformer models and graph neural networks. To make sense of all these different methods, we grouped them into a four-tier classification based on how advanced and how deployable they are. We then studied ten important research papers published between 2015 and 2024, compared them across factors like accuracy, speed, and real-world usability, and identified six gaps that still need to be addressed — including the near-total absence of Indian Sign Language datasets and the difficulty of running these systems on everyday mobile devices. Based on all of this, we outline a practical recognition framework called the Deep Sign Recognition Framework (DSRF) that aims to work in real-world settings, support Indian Sign Language, and run on standard hardware without needing expensive equipment.

Keywords: Sign Language Recognition, Deep Learning, CNN, LSTM, Transformer, Hand Gesture, Computer Vision, Accessibility, Indian Sign Language, Transfer Learning.

I. INTRODUCTION

When we think about communication barriers, we often focus on language differences between countries. But one of the most overlooked barriers exists right within our own communities — the gap between people who are deaf and the hearing world around them. WHO data shows that roughly 430 million people globally experience disabling hearing loss, and by 2050, this figure could exceed 700 million. For many of these individuals, sign language is their first and most natural language. But the vast majority of hearing people have never learned it, which makes even simple everyday interactions unnecessarily difficult.

One of the traditional solutions to this problem has been to rely on human sign language interpreters. While interpreters are valuable, they are not always available, they are expensive to hire, and they cannot be present in every classroom, hospital, or office where a deaf person might need help. This is where technology has an important role to play. Over the past decade, researchers have been working on systems that can automatically recognize sign language gestures from video and convert them into text or speech in real time.

The earliest of these systems used basic image processing tricks — things like detecting skin colour or drawing outlines around the hand. While these methods proved that automated recognition was possible in theory, they broke down very quickly outside of controlled lab settings. A change in lighting, a cluttered background, or a slightly different hand position could be enough to throw off the system completely.

Deep learning changed everything. Convolutional Neural Networks (CNNs) were able to learn what hands actually look like from thousands of example images, without anyone having to define the rules manually. Later, by combining CNNs with recurrent models like LSTMs, researchers could also capture how signs unfold over time — which matters enormously in a language where motion is meaning. More recently, transformer models and Graph Convolutional Networks (GCNs) that work with skeleton joint data have pushed accuracy to record highs on well-known benchmarks like PHOENIX-2014.

Even so, there is still a big distance between what works in a research lab and what can actually help a deaf person on the street. This review paper tries to honestly map out where things stand today. We contribute: (1) a four-tier taxonomy to organize existing SLR systems by their depth and deployability; (2) a structured review of ten representative studies

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154253

spanning 2015 to 2024; (3) a clear comparison of those systems across key dimensions; (4) identification of six real-world gaps that current research has not yet solved; and (5) a conceptual design for the **Deep Sign Recognition Framework (DSRF)** — a system built with practical, real-world use in mind.

II. THEORETICAL BACKGROUND

This section presents the core mathematical models and evaluation metrics that form the technical foundation of deep learning-based sign language recognition systems. Understanding these building blocks is essential for interpreting the design choices and performance results reported in the literature.

A. System Model

An SLR system can be formally defined as a function that maps an ordered sequence of video frames to a predicted sign label or translated output sentence:

$$L = f(V_1, V_2, \ldots, V_t)$$

where each V_i denotes the i-th input video frame, f represents the recognition model encompassing both feature extraction and classification stages, and L is the predicted output — either an isolated sign label or a continuous sentence in the target spoken language. In continuous SLR, the model must handle sequences of signs produced without natural pauses, significantly increasing the complexity of the mapping function.

B. Convolutional Feature Extraction

CNNs form the spatial feature extraction backbone of most SLR systems. The feature map produced by a single convolutional layer is expressed as:

$$F = CNN(V) = ReLU(W * V + b)$$

where V is the input image or video frame, W denotes the learnable convolutional filter weights, b is the bias vector, and * represents the convolution operation. The ReLU activation introduces non-linearity, enabling the network to learn hierarchical representations — from low-level edge and texture features in early layers to high-level semantic hand shape and position features in deeper layers.

C. Temporal Modelling with LSTM

Since sign language is inherently sequential, temporal context across frames is critical for accurate recognition. Long Short-Term Memory (LSTM) networks model this temporal dependency as:

$$h_t = LSTM(F_t, h_{t-1})$$

where F_t is the CNN-extracted feature vector at time step t, h_{t-1} is the hidden state from the previous time step encoding accumulated temporal context, and h_t is the updated hidden state passed to the classifier. This recurrent formulation allows the model to distinguish signs whose hand shapes are similar but whose preceding motion trajectories differ, a capability that purely spatial classifiers lack.

D. Transformer Self-Attention Mechanism

Transformer architectures replace sequential recurrence with a parallel self-attention mechanism that allows every frame in the sequence to directly attend to every other frame:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})V$$

where Q, K, and V are the query, key, and value projections of the input sequence, and d_k is the key vector dimensionality used for scaling. This formulation enables the model to capture long-range temporal dependencies in a single computational step, overcoming the gradient vanishing problem that limits LSTM performance on extended sign sequences. Multi-head attention further allows simultaneous focus on different temporal relationships within the same sequence.

E. Performance Evaluation Metrics

Recognition accuracy for continuous SLR tasks is primarily measured using Word Error Rate (WER), defined as:

$$WER = (S + D + I) / N$$

where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors, and N is the total number of reference sign words. A lower WER reflects fewer recognition errors. For isolated sign classification tasks, standard metrics including classification accuracy and the F1-score are additionally reported to account for class imbalance across the sign vocabulary.

F. System Latency Constraint

For practical accessibility deployment, end-to-end system response time is a critical non-functional requirement. The total recognition latency is modelled as:

$$T\_total = T\_capture + T\_infer + T\_output$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154253

where T_capture is the time required to acquire and pre-process the input video frame, T_infer is the neural network inference time on the target hardware, and T_output is the time to render the recognised sign as text or speech. For a system to be perceived as real-time by a human user, T_total must remain below approximately 200 milliseconds. This constraint directly governs acceptable model size and architecture complexity, particularly for deployment on mobile or embedded edge devices.

III. FOUR-TIER TAXONOMY

To compare sign language recognition systems fairly, you need a way to describe what each one can actually do. We propose grouping them into four tiers based on how sophisticated they are and how ready they are for real-world use.

Tier 1: Rule-Based Systems

These are the oldest type of sign recognition system. They work by defining manual rules — for example, detecting skin-coloured regions in an image, counting fingers, or measuring the size of the hand contour. While they are very lightweight and easy to understand, they fall apart the moment something unexpected happens. Bright sunlight, a colourful background, or a slightly unusual hand position can cause them to fail entirely. They also cannot handle more than a very small vocabulary of signs.

Tier 2: Classical Machine Learning Systems

Tier 2 systems replaced manual rules with trained classifiers. A human expert would still design the features — things like HOG descriptors, optical flow maps, or frequency-domain representations — but a machine learning model like an SVM or Random Forest would learn which combination of features belongs to which sign. These systems are more robust than rule-based ones and can handle larger vocabularies, but they still require significant manual effort and cannot easily generalize to new signing styles or environments.

Tier 3: Deep Learning Classification Systems

From this tier onward, the system learns its own features directly from data. CNNs, LSTMs, GCNs, and transformers all belong here. These models have achieved remarkable accuracy on benchmark datasets and are capable of recognizing hundreds of signs without any hand-crafted features. However, most of them are evaluated in controlled settings and are not designed to run efficiently on everyday hardware. Many also only handle isolated signs rather than the continuous, flowing signing of natural conversation.

Tier 4: End-to-End Deployable Systems (Proposed — DSRF)

This is where we want to get to. A Tier 4 system would take live video from an ordinary camera, run a lightweight recognition model directly on the device, decode continuous sign sequences in real time, and deliver spoken or written output to the hearing person on the other side of the conversation. It would also support Indian Sign Language, not just ASL or German Sign Language. Our proposed Deep Sign Recognition Framework (DSRF) is designed to meet all of these requirements, making it genuinely useful outside of a research lab.

IV. LITERATURE REVIEW

The ten papers reviewed below were chosen because each one represents an important step forward in how we think about and build sign language recognition systems. Together they cover a period from 2015 to 2024 and span isolated sign recognition, continuous sign recognition, multimodal fusion, and transformer-based approaches. Table I gives a full summary.

TABLE I: LITERATURE REVIEW SUMMARY

Sl.	Author(s)	Year & Title	Method / Technique	Key Findings	Venue & Index
1	Pigou L. et al.	2015 – Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition	CNN + Temporal Convolutions	Outperformed earlier pooling methods; strong benchmark for gesture recognition	ChaLearn, Springer
2	Koller O. et al.	2016 – Deep Hand: Training a CNN on 1	CNN with weakly	Showed that large-scale unlabelled data can be leveraged effectively for hand detection	CVPR, IEEE

DOI: 10.17148/IJARCCE.2026.154253

Sl.	Author(s)	Year & Title	Method / Technique	Key Findings	Venue & Index
		Million Hand Images with Weakly Labelled Data	supervised training		
3	Cui R. et al.	2017 – Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition	RCNN with sequence modelling	Reduced segmentation errors in continuous sign streams; better temporal continuity	CVPR, IEEE
4	Camgoz N.C. et al.	2018 – Neural Sign Language Translation	Encoder-Decoder with attention	First end-to-end sign-to-text translation; BLEU-4 score of 18.13 on RWTH-PHOENIX-Weather	CVPR, IEEE
5	Papastratis I. et al.	2021 – Continuous SLR through a Context-Aware Generative Adversarial Network	GAN + CTC decoder	Improved performance across different signers; better temporal alignment of signs	IEEE Access
6	Jiang S. et al.	2021 – Sign Language Recognition with Multi-modal Features	RGB + Depth + Skeleton fusion	96.2% accuracy on DEVISIGN using fused modalities	IEEE Trans. Multimedia
7	Hosain A. et al.	2022 – Hand Gesture Recognition: A Literature Review	Survey of CNN, LSTM, GAN approaches	Reviewed 80+ papers; mapped open problems in real-time gesture-based systems	Int. J. Robotics & Auto.
8	Tunga A. et al.	2022 – Pose-based Sign Language Recognition using GCN	Graph Convolutional Network on skeleton joints	94.3% accuracy; compact model suitable for running on mobile devices	IEEE WACVW
9	Jiang X. et al.	2023 – Skeleton-Aware Multi-Scale Transformer for SLR	Transformer + Skeleton-Aware Module	Achieved 97.5% WER on PHOENIX-2014; strong benchmark result	CVPR, IEEE
10	Bohacek M. & Hruz M.	2024 – Sign Pose-based Transformer for Word-level and Continuous SLR	Pose-based Vision Transformer (ViT)	Top performance on multiple benchmarks; capable of running in real time	WACV, IEEE

V. COMPARATIVE ANALYSIS

Looking at these ten studies together, some clear patterns emerge. The earliest CNN-based systems proved that automated sign recognition is feasible, but they handled only isolated signs in constrained settings. The moment you need a system to deal with signs flowing together in natural conversation, you need something more than a frame-by-frame classifier — you need a model that understands sequence.

The papers by Cui et al. and Camgoz et al. were important turning points. Cui et al. showed that combining CNNs with recurrent models could handle continuous signing without needing manual segmentation. Camgoz et al. went further and built the first end-to-end system that could translate a signing video directly into a grammatically correct spoken language sentence — treating sign language translation the same way you would treat machine translation between two spoken languages.

The multimodal approach by Jiang et al. (2021) is interesting because it shows what is possible when you can use multiple types of input simultaneously. By combining regular video, depth images, and skeleton joint positions, they reached

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154253

96.2% accuracy on DEVISIGN. The downside is that you need a depth camera to make it work, which most ordinary people do not have.

Skeleton-based approaches like Tunga et al. and the transformer systems like Jiang et al. (2023) and Bohacek and Hruz offer a more practical path forward. By working with skeleton keypoints rather than raw pixels, these models are lighter and more generalizable. The skeleton captures the geometry of what the hands and body are doing without being distracted by background clutter or clothing colour. The best of these models now achieve near-perfect accuracy on established benchmarks.

That said, almost every one of these papers shares the same blind spot: they were all built and tested on Western sign language datasets. Indian Sign Language, despite being used by millions of people, does not have a single large-scale, publicly available benchmark dataset. This means the methods that work so well on PHOENIX-2014 or ASL datasets have never been seriously tested in an Indian context.

TABLE II: COMPARATIVE ANALYSIS OF REVIEWED SYSTEMS

Sl.	Paper	Technique	Performance	Advantages	Limitations
1	Pigou et al.	CNN + Temporal Conv	High	Captures hand motion over time	No continuous sign support
2	Koller et al.	Weakly supervised CNN	High	Works with large unlabelled data	Needs big data to train well
3	Cui et al.	RCNN + Sequence model	High	Handles unbroken signing streams	Heavy on compute resources
4	Camgoz et al.	NMT Encoder-Decoder	BLEU-4: 18.13	Translates signs to full sentences	Needs paired sign-text corpora
5	Papastratis et al.	GAN + CTC	High	Works across different signers	Training pipeline is complex
6	Jiang S. et al.	RGB+Depth+Skeleton	96.2%	Uses rich combined input	Needs a depth camera
7	Hosain et al.	Survey only	N/A	Covers a wide range of methods	No new system was built
8	Tunga et al.	GCN on skeleton	94.3%	Lightweight; runs on phones	Skeleton detection adds delay
9	Jiang X. et al.	Transformer + Skeleton	97.5% WER	Best-in-class benchmark accuracy	Uses a lot of memory
10	Bohacek & Hruz	Pose-based ViT	Top benchmark	Fast enough for real-time use	Still relies on pose estimation

VI. RESEARCH GAP

Even with all the impressive progress reviewed above, there are some important problems that the field has not yet solved. We identified six gaps that are worth paying close attention to.

Gap 1 — No Large-Scale Indian Sign Language Dataset Exists: This is arguably the most pressing gap for researchers in India. Without a dataset comparable to PHOENIX-2014 for ISL, training and evaluating models for the Indian deaf community is extremely difficult. The absence of this resource means that Indian Sign Language is effectively invisible in most of the research literature.

Gap 2 — Systems Still Struggle with Unseen Signers: Most published accuracy numbers come from experiments where the same signers appear in both the training and test data. When you test these systems on a person they have never seen

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154253

before, accuracy drops significantly. A truly useful system needs to work for anyone, regardless of their hand size, skin tone, signing speed, or regional signing variation.

Gap 3 — High-Accuracy Models Cannot Run on Ordinary Devices: The models that score best on benchmarks — especially transformer-based ones — are too large and slow to run on a smartphone or a small embedded device without significant engineering work. Techniques like model pruning, quantization, and knowledge distillation exist, but applying them specifically to SLR models is still very much an open research problem.

Gap 4 — Facial Expressions Are Being Ignored: In sign language, facial expressions are not just emotional — they are grammatical. Raised eyebrows can turn a statement into a question. Mouth movements can distinguish between signs that look identical with the hands. Nearly every system reviewed here focused only on hand gestures, which means they are missing a critical part of the language.

Gap 5 — Communication Only Works in One Direction: Current SLR systems translate signs into text or speech, but they do not help a hearing person communicate back to a deaf person in sign language. For a real two-way conversation, you also need sign language synthesis — a system that takes spoken words and generates the corresponding signing animation or video.

Gap 6 — Real-World Testing Is Extremely Rare: Almost everything in the literature was evaluated in a controlled lab environment with a plain background, good lighting, and a single signer in frame. The real world is messier than that. We simply do not know how well these systems would hold up in a crowded street, an outdoor setting with variable lighting, or a video call with a shifting camera angle.

VII.CONCLUSION

Sign language recognition has come a remarkably long way in a short time. What started as fragile, rule-based systems that could only recognize a handful of signs under perfect conditions has evolved into transformer and GCN-based models that achieve near-human accuracy on established benchmarks. That progression is genuinely impressive, and it reflects how much deep learning has transformed computer vision as a whole.

But the gap between what works in a research paper and what can actually help a deaf person in everyday life is still significant. The field needs better datasets — especially for Indian Sign Language. It needs models that generalize to new signers without needing to be retrained. It needs systems that are lightweight enough to run on a phone. And it needs to treat sign language as a complete linguistic system, not just a collection of hand shapes.

The Deep Sign Recognition Framework (DSRF) we propose in this review is our attempt to point toward what that kind of system could look like. By combining a lightweight transformer-based recognition model, continuous sign decoding with CTC, non-manual marker integration, support for Indian Sign Language, and deployment on standard consumer hardware, DSRF is designed to be useful in the real world rather than just impressive on a benchmark.

There is still a lot of work to do. Building an Indian Sign Language dataset will require collaboration with the deaf community. Making models robust to new signers and real-world conditions will take new training strategies. But none of these are impossible problems. With the right research priorities and genuine involvement of the deaf community in shaping these systems, sign language recognition can become a tool that truly makes a difference in people's daily lives.

REFERENCES

[1] L. Pigou, M. Van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," International Journal of Computer Vision, vol. 126, no. 2, pp. 430–439, 2018.

[2] 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.

[3] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in Proc. IEEE CVPR, 2017, pp. 4141–4150.

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

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154253

[5] I. Papastratis et al., "Continuous sign language recognition through a context-aware generative adversarial network," Sensors, vol. 21, no. 7, p. 2437, 2021.

[6] S. Jiang et al., "Skeleton aware multi-modal sign language recognition," in Proc. IEEE CVPRW, 2021, pp. 3413–3423.

[7] A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Kosecka, "Hand gesture recognition: A literature review," International Journal of Robotics and Automation, vol. 37, no. 1, 2022.

[8] A. Tunga, S. Nuthalapati, and J. Wachs, "Pose-based sign language recognition using GCN and BERT," in Proc. IEEE WACVW, 2021, pp. 31–40.

[9] X. Jiang, T. Takiguchi, and Y. Ariki, "Skeleton-aware multi-scale transformer for sign language recognition," in Proc. IEEE CVPR, 2023, pp. 21453–21463.

[10] M. Bohacek and M. Hruz, "Sign pose-based transformer for word-level and continuous sign language recognition," in Proc. IEEE WACV, 2022, pp. 3314–3323.

