

Deep Learning for Indian Sign Language Processing: A Survey of Datasets, Methodologies, and Future Prospects

Ravindra. B. Murumkar¹, Pradyumna Wagholikar², Kaustubh Pawar³, Ved Pingle⁴, Pallavi Ranamale⁵

¹ Professor, Dept. of IT, Pune Institute of Computer Technology, Pune, Maharashtra, India.

^{2,3,4,5} Student, Dept. of IT, Pune Institute of Computer Technology, Pune, Maharashtra, India.

Abstract - Indian Sign Language (ISL) is the primary means of communication for millions of individuals within India's Deaf and Hard of Hearing (DHH) community. However, a significant communication gap persists due to the scarcity of certified human interpreters, motivating the development of automated ISL recognition and translation systems. This survey presents a comprehensive review of the technical evolution of ISL processing, tracing the progression from early computer vision methods based on handcrafted features to modern deep learning paradigms. Convolutional Neural Networks (CNNs) have advanced spatial feature extraction, while Recurrent Neural Networks (LSTMs, GRUs) enhanced temporal modeling of sequential gestures. More recently, Transformer architectures have enabled end-to-end translation with improved contextual understanding. A prominent two-stage paradigm has emerged—combining pose estimation frameworks such as MediaPipe for skeletal keypoint extraction with deep learning models for dynamic sequence interpretation. The survey also emphasizes the growing importance of datasets, highlighting the transition from small, custom-collected corpora to large-scale public benchmarks like CISLR, ISLTranslate, and iSign. Despite notable progress, challenges remain, including limited data standardization, occlusions in two-handed signs, and insufficient modeling of non-manual markers. The paper concludes by outlining future research directions, emphasizing large-scale unified datasets, multimodal integration, and advanced Transformer-based architectures for robust, real-time ISL understanding.

Key Words: Indian Sign Language, Deep Learning, Computer Vision, Sign Language Recognition, Sign Language Translation, Pose Estimation, MediaPipe, CNN, LSTM, GRU, Transformers.

1. INTRODUCTION

Language is the cornerstone of human connection and societal participation. In India alone, approximately 63 million people rely on Indian Sign Language (ISL) as their primary means of communication [1], [2], [5], [8]. ISL is a complete and complex visual-gestural language—far more than a collection of gestures—with its own grammar, syntax, and linguistic structure distinct from the spoken languages of the region [3]. According to Ethnologue, ISL ranks among

the most widely used sign languages globally, underscoring its social and cultural importance [4].

Despite its widespread use, a profound communication gap persists between the Deaf and Hard-of-Hearing (DHH) community and the hearing population. This gap is largely due to a critical shortage of qualified human interpreters. Official estimates suggest that there are only around 300 certified ISL interpreters in India, a number vastly insufficient to serve a community of millions [2], [4], [5]. This deficit has far-reaching consequences, creating barriers to essential services such as healthcare, education, and legal support [1], [5]. In these critical contexts, the absence of effective communication can result in misdiagnoses, educational disadvantages, and inequitable legal outcomes.

The urgent need to bridge this gap provides strong motivation for developing automated, technology-driven solutions for ISL recognition and translation. Advances in computer vision and deep learning have opened new possibilities for such systems, which hold the potential to democratize access to information and foster inclusion for the DHH community [1], [3].

However, building robust automated systems for ISL is a formidable task due to several intrinsic linguistic and computational challenges. Unlike American Sign Language (ASL), which is predominantly one-handed, ISL is fundamentally a two-handed language, with many signs requiring coordinated hand movements [3], [11]. This introduces the persistent problem of occlusion, where one hand can obscure the other, leading to a loss of critical visual information for recognition systems [3], [10].

Furthermore, ISL communication is inherently multi-modal. Non-Manual Markers (NMMs)—such as facial expressions, head movements, and body posture—serve as essential grammatical elements that modify or define the meaning of signs [5], [7], [10]. The absence of a single standardized version of ISL has also led to significant regional variations, creating a major generalization challenge for data-driven models [3].

To effectively address these challenges, it is essential to clearly define the core computational tasks in automated ISL processing. The two principal tasks are Sign Language Recognition (SLR) and Sign Language Translation (SLT). SLR

focuses on identifying and classifying individual signs or sequences of signs, typically producing a sequence of glosses (written labels representing signs) [3], [11]. SLT, on the other hand, seeks to translate continuous sequences of signs into grammatically correct sentences in a target spoken language such as English. This represents a true machine translation problem, as the syntactic structures of ISL and spoken languages differ fundamentally [3], [11].

The evolution of ISL recognition research reflects a significant methodological progression. Early systems relied on traditional computer vision and image processing techniques. For instance, a 2011 study by Rajam and Balakrishnan proposed a rule-based system for recognizing 32 signs using binary finger position analysis combined with Canny Edge Detection and handcrafted feature extraction [6]. While innovative for their time, such approaches were brittle, computationally intensive, and highly dependent on controlled conditions.

The advent of deep learning introduced a paradigm shift—moving from handcrafted features to hierarchical feature learning directly from raw data. Convolutional Neural Networks (CNNs) demonstrated strong capabilities for spatial feature extraction, while later methods incorporated pose estimation with Recurrent Neural Networks (RNNs) to model temporal dependencies [1]. This transition has greatly enhanced the robustness and scalability of modern ISL recognition systems.

This paper presents a comprehensive survey of the progress in automated ISL processing. We review the critical evolution of datasets that have driven the field's growth, trace the methodological trajectory from traditional vision-based techniques to state-of-the-art deep learning architectures, and analyze the transformative impact of enabling technologies such as pose estimation. Furthermore, we identify key challenges, highlight performance gaps, and outline future research directions aimed at developing linguistically grounded, inclusive, and real-world-deployable ISL recognition and translation systems.

2. THE EVOLUTION OF ISL DATASETS

The advancement of automated ISL processing is inextricably linked to the availability and quality of data resources. For years, the field was hampered by a “data desert,” which limited research to small-scale, often incomparable studies. The recent emergence of large, public benchmarks has been the single most important catalyst for progress, enabling the application of data-hungry deep learning models.

Early research was characterized by the use of small, bespoke datasets created by individual research groups [3], [6]. These datasets typically featured a limited vocabulary, a small number of signers, and were recorded in highly controlled laboratory environments to simplify the

computer vision task [3], [6]. While a necessary first step, this practice led to models that lacked real-world generalizability and made it difficult to compare different approaches meaningfully [4].

A turning point for the field was the release of the first large-scale, publicly available datasets. The INCLUDE dataset, for instance, provided over 4,000 videos for 263 common word signs, offering a more substantial resource for word-level recognition [1], [3]. This was followed by the CISLR (Corpus for Indian Sign Language Recognition) dataset, which represented a major leap forward with a vocabulary of approximately 4,700 words across 7,050 videos [4]. Its unique structure, with a low average of 1.5 videos per word, spurred research into one-shot learning paradigms, where models must learn to recognize a sign from a single example [4].

The focus of the research community shifted from recognition to translation with the publication of ISLTranslate. As the first large-scale translation dataset for ISL, it contained approximately 31,000 parallel ISL-English sentence and phrase pairs, enabling the training of modern end-to-end translation architectures for the first time [2], [10].

The current state of the art in ISL data resources is represented by the iSign benchmark, released in 2024 [10]. By consolidating previous major datasets like CISLR and ISLTranslate and augmenting them with new data from authentic sources such as ISLRTC and ISH News, iSign provides a massive resource of over 118,000 video-sentence pairs [10].

Crucially, iSign is a multi-task framework that proposes a standardized set of challenges, including SignVideo2Text Translation, Text2Pose Generation, and Sign Semantic Similarity, thereby guiding the research community toward more holistic and capable systems [10]. The “benchmark effect” of these datasets has been profound, actively shaping the research agenda from isolated sign classification toward the more ambitious goal of continuous, multi-modal sign language understanding and generation.

3. METHODOLOGICAL TRAJECTORIES IN ISL PROCESSING

The evolution of methodologies for ISL processing mirrors broader trends in computer vision and natural language processing, marked by a progression from handcrafted features to sophisticated, end-to-end deep learning architectures.

A. Early Approaches: Handcrafted Features

Initial forays into SLR were rooted in traditional computer vision paradigms. These systems typically involved a two-stage process: first, extracting handcrafted visual features from images using descriptors like Scale-Invariant Feature

Transform (SIFT), Histogram of Oriented Gradients (HOG), or contour analysis [3], [6], [7]; and second, feeding these feature vectors into classical machine learning classifiers such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), or Hidden Markov Models (HMM) for recognition [3], [5], [7].

While foundational, these methods were often brittle and highly sensitive to variations in lighting, background, and signer appearance [4], [5].

B. The Rise of Deep Learning: CNNs, LSTMs, and Transformers.

The advent of deep learning triggered a paradigm shift by enabling the automatic learning of feature hierarchies directly from data.

- Convolutional Neural Networks (CNNs): CNNs became the standard for spatial feature extraction, first applied to static sign recognition by treating it as an image classification problem [3], [7]. Architectures like InceptionV3 and custom CNNs achieved high accuracy on constrained tasks like alphabet recognition [3]. For dynamic signs, 3D CNNs were introduced to learn spatio-temporal features directly from video volumes, though at a high computational cost [7].

- Recurrent Architectures: To explicitly model the temporal dynamics of sign language, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, were adopted [1], [5], [8], [9]. This led to a dominant hybrid architecture where a CNN extracts spatial features from each video frame, and an LSTM or GRU processes the resulting sequence of features to model temporal dependencies [7], [9]. Attention mechanisms were often added to allow the model to focus on the most salient frames in a sequence [7].

- Transformers: The current state of the art in sequence-to-sequence tasks is the Transformer architecture, whose self-attention mechanism allows it to model long-range dependencies more effectively than RNNs [10], [11]. A key breakthrough was the development of an architecture that jointly learns recognition and translation in a single end-to-end model [11]. In this framework, a Connectionist Temporal Classification (CTC) loss is applied to the encoder's output to supervise the recognition of sign glosses.

This intermediate supervision forces the model to learn a meaningful sign representation, which is then used by the decoder for the final translation task, significantly improving performance over direct video-to-text models [11].

C. The Role of Pose Estimation Frameworks

A critical enabling technology that has reshaped the field is real-time pose estimation. Frameworks like Google's MediaPipe Holistic can process a video frame and output a structured, skeletal representation comprising 3D

coordinates for keypoints on the hands (21 per hand), body (33), and face (468) [5], [8], [9]. This transforms the input from a high-dimensional pixel space to a low-dimensional, semantically rich feature space.

This approach offers several profound advantages. It is computationally efficient, making real-time applications on consumer hardware feasible [1], [5]. The skeletal data is inherently more robust to variations in lighting, background, and clothing [1]. It also helps mitigate privacy concerns by anonymizing the signer's appearance [9].

This pipeline effectively decouples the computer vision problem from the sequence modeling task, allowing researchers to leverage a powerful, off-the-shelf feature extractor and focus on designing sophisticated sequential models [5], [8], [9]. Consequently, many state-of-the-art systems now follow this pipeline: video is fed into MediaPipe, and the resulting sequence of landmark vectors is processed by an LSTM, GRU, or Transformer network, often achieving recognition accuracies exceeding 95% [1], [5], [8].

4. COMPARATIVE ANALYSIS AND PERFORMANCE

This section presents a comparative evaluation of major approaches developed for Indian Sign Language (ISL) recognition and translation. The analysis highlights performance trends, methodological advancements, and the datasets that have driven progress in this domain.

The progression of research in ISL recognition is marked by increasing methodological sophistication and performance, largely driven by the availability of better datasets and more powerful computational models. A comparative analysis of key studies, as summarized in Table ??, reveals clear trends in the field.

Early works, such as that by Rajam et al. [6], relied on traditional image processing techniques like feature point extraction on small, custom datasets of static signs. While achieving high accuracy (98.1%) within their constrained, single-user environment, these methods lacked the ability to generalize to new signers or dynamic gestures.

The adoption of deep learning, particularly CNNs, marked a significant step forward. Studies like Sharma et al. [3] focused on CNN-based recognition of static alphabets, demonstrating the power of learned features but also highlighting the limitations of existing datasets in terms of diversity and the challenge of two-handed signs.

A major breakthrough came with the integration of pose estimation frameworks and recurrent neural networks to handle dynamic signs. The work of Shetty et al. [1] and Rawat et al. [5], both leveraging MediaPipe for feature extraction and LSTMs for sequence modeling, achieved

impressive accuracies of 98% and 96.97%, respectively, on isolated dynamic signs. These studies underscore the effectiveness of the pose-estimation pipeline for creating robust, real-time systems that are invariant to background and lighting conditions.

Subramanian et al. further refined this approach by proposing an optimized GRU (MOPGRU) model, which demonstrated faster convergence and higher learning efficiency [9].

The focus has recently expanded from recognition to translation, a shift enabled by new datasets and architectures. The work by Camgoz et al. on the Transformer

architecture, while not on ISL, introduced a seminal joint recognition-translation framework that has become highly influential [11].

The creation of the ISLTranslate [2] and iSign [10] datasets has now provided the necessary resources to apply such advanced translation models to ISL, setting the stage for the next generation of research in the field.

TABLE I: Comparative Analysis of Representative ISL Recognition and Translation Studies

Paper	Core Methodology	Dataset (reported)	Task / Protocol	Metric	Strengths	Weaknesses / Limitations
Rajam et al. (2011) [6]	Feature-point extraction and image processing	Custom (32 static signs)	Isolated static-sign recognition; single user	98.1% Accuracy (isolated, controlled)	High accuracy under controlled conditions; early ISL work	Single signer, static-only signs; limited generalization to continuous signing
Sharma et al. (2025) [3]	Convolutional Neural Network (CNN)	Custom (2,600+ static images)	Letter/number/static-sign recognition	High accuracy on letters and numbers	Addresses two-handed signs; optimized for affordable hardware	Primarily static signs; limited cross-paper comparability due to custom dataset
Shetty et al. (2024) [1]	MediaPipe pose estimation with LSTM	INCLUDE (as per paper)	Isolated word/sign recognition; real-time setting	98% Accuracy (isolated signs)	Real-time performance; robust to clothing and background variations	Evaluated on isolated signs; not demonstrated for continuous translation
Subramanian et al. (2022) [9]	MediaPipe with optimized GRU (MOPGRU)	Custom + WLASL, LSA64	Small-gesture recognition; signer-dependent and signer-independent splits	95% Accuracy (task-dependent)	Faster convergence and improved learning efficiency	Mostly small datasets; limited vocabulary for large-scale generalization
Rawat et al. (2025) [5]	MediaPipe Holistic with sequential LSTM	Custom (11 dynamic and static gestures)	Isolated dynamic/static gesture recognition; signer-independent evaluation	96.97% Accuracy (isolated gestures)	Robust across lighting conditions; signer-independent evaluation	Very small dataset; not evaluated on sentence-level translation
Joshi et al. (2022) [4]	One-shot learning using ASL-pretrained I3D features	CISLR (4,700 words, 7,050 videos)	Large-vocabulary word-level recognition; one-shot evaluation	16.8% Top-1 Accuracy	Introduced large-vocabulary ISL dataset; explored cross-lingual transfer	Low absolute accuracy; highlights challenge of one-shot learning in ISL
Camgoz et al.	Sign Language	PHOENIX14T	Sentence-level sign	~2× BLEU-4	Strong end-to-	Evaluated on

(2020) [11]	Transformer (joint recognition and translation using CTC + seq2seq)	(German SL)	language translation (video-to-text)	improvement over previous baselines	end architecture; introduced intermediate CTC supervision	German SL, not ISL; cross-lingual generalization uncertain
Joshi et al. (2024) [10]	Multi-task benchmark creation (iSign): translation, generation, similarity	iSign (118k pairs)	Multi-task benchmark: SignVideo2Text, Text2Pose, semantic-similarity	Baseline BLEU-4 score of 1.47	Largest consolidated ISL resource; establishes standard tasks and evaluation splits	Baseline performance low, indicating scope for improved modeling

In summary, the comparative analysis highlights a clear paradigm shift from handcrafted and pixel-based recognition techniques toward pose-driven, deep learning architectures, culminating in end-to-end Transformer-based translation frameworks empowered by large-scale datasets like ISLTranslate and iSign.

5. PREVAILING CHALLENGES AND LIMITATIONS

Despite significant progress, the field of automated ISL processing faces several persistent challenges that define the frontier of current research. These limitations are frequently cited across the literature and must be addressed to develop truly practical systems.

A primary challenge remains the scarcity and diversity of data. While benchmarks like iSign represent a monumental step forward, the volume of data is still orders of magnitude smaller than that available for spoken languages [10]. Furthermore, existing datasets often lack sufficient diversity in terms of signers, regional dialects, and recording environments, which can lead to models that do not generalize well to real-world “in-the-wild” conditions [3].

The inherent linguistic complexity of ISL poses another major hurdle.

- **Two-Handed Signs and Occlusion:** The frequent use of two-handed gestures in ISL leads to persistent occlusion, where one hand blocks the other from view. This remains a difficult computer vision problem that can cause the loss of crucial handshape information [3], [10], [11].
- **Non-Manual Markers (NMMs):** The grammatical information conveyed through facial expressions, head tilts, and body posture is critical for accurate interpretation. While pose estimation frameworks can extract facial landmarks, current models struggle to effectively integrate these NMMs in a linguistically meaningful way [5], [7], [10].
- **Fingerspelling and Role Shifts:** Signers often use fingerspelling for names or technical terms, requiring models to switch from word-level to character-level

recognition. Additionally, role shifts, where a signer embodies a character in a narrative, are complex phenomena that are not well-handled by current architectures [10].

There is also a significant “lab-to-life” performance gap. Models that achieve high accuracy on clean benchmark data often experience a sharp performance degradation when deployed in real-world scenarios with cluttered backgrounds, variable lighting, and novel signers [1].

Finally, the community faces a challenge in evaluation. Standard NLP metrics like BLEU, designed for linear text, may be inadequate for assessing the quality of translation for a visual-spatial language like ISL, as they fail to capture the preservation of spatial grammar or other visual linguistic features [10].

6. FUTURE RESEARCH DIRECTIONS

Addressing the prevailing challenges in ISL processing requires a multi-pronged research effort focused on data, models, and evaluation. Based on the limitations identified in the existing literature, several key directions for future work emerge.

First, there is a continued need for dataset expansion and enrichment. Future data collection efforts should prioritize capturing a wider diversity of signers from different regions to better represent ISL’s dialectal variations. Datasets should also include more continuous, conversational signing recorded in naturalistic “in-the-wild” environments. Crucially, annotations should be expanded to include explicit labels for non-manual markers and other linguistic phenomena to facilitate the development of more nuanced models.

Second, the field must move toward linguistically-informed model architectures. Rather than relying on generic sequence-to-sequence models, future research should explore architectures that explicitly incorporate the unique properties of ISL. This could involve using graph neural networks to model the dynamic spatial relationships between body parts or designing specialized attention

mechanisms that are aware of the signing space and the distinct roles of the dominant and non-dominant hands. Developing more sophisticated multi-modal fusion techniques to effectively integrate manual and non-manual channels is a critical priority.

Third, given the persistent data scarcity, self-supervised and few-shot learning techniques are vital. The success of using a model pre-trained on ASL for one-shot recognition in ISL suggests that cross-lingual transfer learning is a promising avenue [4]. Pre-training models on large, unlabeled sign language video corpora could enable the learning of powerful, generalizable representations.

Finally, as the technology matures, research must increasingly focus on human-centered evaluation and community engagement. This involves developing new evaluation metrics that better correlate with human judgments of translation quality and are sensitive to the linguistic nuances of ISL [10].

More importantly, it requires a shift from designing systems for the Deaf community to co-designing them with the community, ensuring that the resulting technology is not only accurate but also culturally appropriate, respectful, and genuinely useful to its intended users.

6. CONCLUSION

This survey has charted the rapid evolution of automated Indian Sign Language processing, a field that has transitioned from foundational computer vision exercises to the forefront of multi-modal deep learning research.

The journey has been characterized by a methodical progression from recognizing isolated, static signs to the ambitious goal of translating continuous, dynamic sign language. This advancement has been fundamentally driven by two parallel forces: the development of increasingly sophisticated neural architectures—from CNNs and LSTMs to Transformers—and the critical emergence of large-scale, public datasets like CISLR, ISLTranslate, and the comprehensive iSign benchmark.

The current state of the art is defined by end-to-end systems that leverage pose estimation frameworks like MediaPipe for efficient and robust feature extraction, coupled with Transformer-based models that can jointly learn to recognize and translate signs.

Despite achieving impressive performance on benchmark tasks, significant challenges remain. The linguistic complexities of ISL, including the prevalence of two-handed signs, the grammatical importance of non-manual markers, and the use of spatial grammar, continue to test the limits of current models.

Overcoming these hurdles and closing the gap between benchmark performance and real-world reliability will define the next phase of research. Future work must focus on creating larger and more diverse datasets, designing linguistically-informed architectures, and engaging deeply with the Deaf community to ensure the development of technology that is both powerful and purposeful.

The continued progress in this domain holds immense potential to dismantle communication barriers and foster a more inclusive and accessible society.

REFERENCES

- [1] S. Shetty, E. Hirani, A. Singh, and R. Koshy, "Gesture-to-Text: A Real-Time Indian Sign Language Translator with Pose Estimation and LSTMs," *Procedia Computer Science*, vol. 235, pp. 2684-2692, 2024.
- [2] A. Joshi, S. Agrawal, and A. Modi, "ISLTranslate: Dataset for Translating Indian Sign Language," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.
- [3] N. Sharma, J. Mandal, A. Chaudhury, and S. V., "Performance Analysis of CNN-Based Indian Sign Language Recognition," in *Proceedings of the International Conference on Advanced Research in Electronics and Communication Systems (ICARECS 2025)*, Atlantis Highlights in Engineering, vol. 38, 2025, pp. 15-24.
- [4] A. Joshi et al., "CISLR: Corpus for Indian Sign Language Recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 10357-10366.
- [5] P. Rawat, P. Kumar, V. K. Tamta, and A. Kumar, "A Comprehensive Approach to Indian Sign Language Recognition: Leveraging LSTM and MediaPipe Holistic for Dynamic and Static Hand Gesture Recognition," *EAI Endorsed Transactions on AI and Robotics*, vol. 4, 2025.
- [6] P. S. Rajam and G. Balakrishnan, "Real time Indian Sign Language Recognition System to aid deaf-dumb people," in *2011 IEEE 13th International Conference on Communication Technology*, 2011, pp. 737-742.
- [7] A. Singh, A. Wadhawan, M. Rakhra, U. Mittal, A. Al Ahdal, and S. K. Jha, "Indian Sign Language Recognition System for Dynamic Signs," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2022.
- [8] R. Ba, R. Holla, A. Azam, and P. S. Kiran, "Sign language recognition using LSTM," *SSRN Electronic Journal*, 2024.

[9] B. Subramanian et al., "An integrated mediapipe-optimized GRU model for Indian sign language recognition," *Scientific Reports*, vol. 12, no. 1, p. 11964, 2022.

[10] A. Joshi et al., "iSign: A Benchmark for Indian Sign Language Processing," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[11] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10023-10033.