

AI-Based Sign Language Recognition and Conversion to Text Using Computer Vision: A Review

Tejas Shantaram Jadhav, Prof. Bangar A.P., Prof. Bhosale S.B., Dr. A. A. Khatri

Department of Computer Engineering
Jaihind College of Engineering, Kuran, Pune, India

Abstract- This paper reviews recent computer-vision approaches that convert sign language into text. Sign language is the main means of communication for people with hearing and speech impairments, yet most hearing people cannot understand it. Four recent IEEE Access papers are studied: a comprehensive survey of traditional and deep-learning recognition methods, a two-stream graph convolution network with attention and residual connections, a Vision Transformer for static Indian signs, and a lightweight pose-based graph convolution pipeline called SignGraph. The review compares their input modalities, models, datasets, accuracy and computational cost. The study shows that skeleton (pose) based methods built on MediaPipe-style landmarks give accuracy close to or better than RGB-based models while needing about twenty times fewer parameters, which suits real-time webcam use. Open problems such as large vocabularies, signer independence, non-manual features and the scarcity of Indian Sign Language data are discussed, and directions for future work are suggested.

Keywords— Sign language recognition, computer vision, deep learning, graph convolutional network, vision transformer,

I. INTRODUCTION

Sign language is a visual language in which meaning is carried by hand shape, position, movement, body posture and facial expression. According to the World Health Organization, about 430 million people need rehabilitation for disabling hearing loss, and the number of people with some degree of hearing loss is projected to reach about 2.5 billion by 2050 [1], [5]. Most hearing people do not know sign language, so deaf and hard-of-hearing people face a communication barrier in hospitals, banks, schools and public offices. Human interpreters are limited in number and costly, which motivates automatic sign language recognition (SLR).

The core task of vision-based SLR is to take video of a signer, usually from an ordinary camera, and output the corresponding text [1]. The problem touches several research areas: hand detection, pose estimation, action recognition and sequence modelling. Early systems relied on sensor gloves or coloured gloves. Image-processing methods removed the glove but used hand-crafted features, and deep learning later allowed features to be learned automatically from data.

The aim of this paper is to review four recent papers from IEEE Access that represent different directions in SLR: a comprehensive survey [1], a two-stream graph convolution model [2], a Vision Transformer for static Indian signs [3] and a lightweight pose-based graph model [4]. Section II gives background, Section III summarises each paper, Section IV describes the common sign-to-text pipeline, Section V covers datasets and metrics, Section VI compares the methods, and Sections VII to IX present challenges, future scope and conclusion.

II. BACKGROUND

1. Types of Sign Language Recognition

Signs can be static or dynamic. A static sign, such as a fingerspelled letter, is fully described by a single image, while a dynamic sign needs a sequence of frames [2], [3]. In video, the basic unit is the gloss, a complete sign word. Isolated SLR (ISLR) recognises one gloss per video, whereas continuous SLR (CSLR) recognises a sentence of several glosses and is a sequence-to-sequence problem [1]. Sign languages also differ between countries, for example American (ASL), British (BSL) and Indian (ISL) sign language, which makes a universal model difficult [2].

2. Traditional Methods

Traditional SLR splits the problem into feature extraction and temporal modelling [1]. Features came from data gloves with attached sensors, from coloured gloves tracked by a camera, or from image-processing operators such as HOG, SIFT and SURF. Temporal models such as Hidden Markov Models (HMM), Conditional Random Fields (CRF) and Dynamic Time Warping (DTW) then mapped the features to words. These methods were interpretable but sensitive to illumination and occlusion, needed costly manual feature design, and were mostly tested on small datasets [1].

3. Deep Learning Methods

Deep learning methods appeared after 2012 and learn features directly from data [1]. Common designs are 2D CNNs combined with RNN, LSTM or temporal convolution networks, 3D CNNs such as C3D and I3D, Transformers, generative adversarial networks (GAN) and graph convolution networks (GCN) [1]. Inputs can be RGB frames, depth, or skeleton key points. RGB is the most common because it can be captured by a phone or webcam, while skeleton data removes background and clothing and so improves robustness to lighting and scene changes [1], [2].

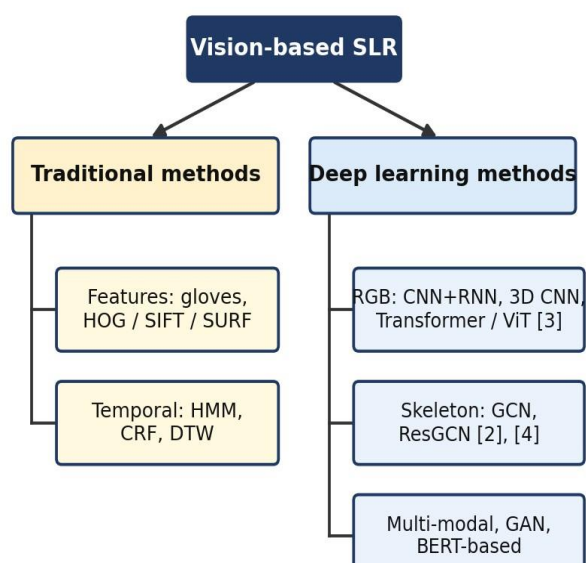


Fig.1. Taxonomy of vision-based sign language recognition approaches

III. REVIEW OF SELECTED PAPERS

1. Comprehensive Review of Traditional and Deep Learning Approaches [1]

Tao et al. survey SLR from early glove-based systems to Transformers. Their main idea is to describe every method through two parts, feature extraction and temporal modelling, and to compare traditional and deep approaches on that basis. They cover evaluation metrics (WER, BLEU and ROUGE), more than a dozen public datasets, and tables of CNN/RNN, Transformer and other methods with their inputs and results. They note that isolated SLR is near perfect on small datasets such as CSL but reaches only about 57% on the larger WLASL dataset, that Transformer-based methods usually bring WER below 25 on continuous benchmarks, and that most work is not tested for unseen signers or unseen sentences. They list six open problems: short training videos, signer-independent and unseen-sentence recognition, signs outside the dataset, multi-person scenes, model complexity and online recognition.

2. Two-Stream Graph Convolution with Attention (GCAR) [2]

Miah et al. use 67 whole-body key points (42 from the hands, 5 from the body and 20 from the face) instead of hand points only, because body motion and facial expression add information for large vocabularies. The model has two streams. One takes the joint skeleton and the other takes joint motion, computed as the coordinate difference between consecutive frames. Each stream uses a separable temporal convolution network (Sep-TCN) and stacked graph convolution blocks with channel attention and residual connections, and the two features are fused for classification. The model has only 0.69 million parameters. It reaches 63.25% top-1 and 90.31% top-10 accuracy on WLASL-100, about 94% on the Pakistani Sign Language dataset, 99.75% on the Mexican Sign Language dataset and 34.41% on the 2,745-class ASLLVD dataset. Their ablation shows that attention consistently improves accuracy.

3. SignFormer: Vision Transformer for Static Indian Signs [3]

Kothadiya et al. apply a Vision Transformer encoder to static Indian Sign Language. Images are resized to 72×72 , split into 6×6 patches (144 patches), given positional embeddings, passed through multi-head self-attention layers and classified by an MLP head. The dataset has 36 classes (digits 0 to 9 and the alphabet). With flip, rotation and colour augmentation the model reaches 99.29% accuracy after only five epochs, and accuracy rises from 97.91% with one attention head to 99.29% with four heads. Results on American (98.31%) and Bangla (98.79%) static sign datasets show good generalisation. The work is limited to static signs and does not model motion.

Naz et al. extract poses with MediaPipe Holistic and keep 65 joints (21 per hand, 11 head and 12 upper-body joints) using only 2D (x, y) coordinates, because depth values are less reliable. Each video is sampled to 64 frames and fed to a ResGCN network with bottleneck blocks and residual connections. SignGraph gives top-1 accuracy of 72.09%, 71.40% and 61.83% on WLASL-100, 300 and 1000 and 100% on the LSA-64 dataset. It outperforms the earlier pose-based methods and compares well with the RGB-based I3D model while using 0.62 million parameters against 12.4 million, 1.95 against 5.22 GFLOPs, and 0.040 s against 0.50 s inference time per video. Error analysis shows that about 43% of errors come from different words signed in the same way and about 34% from signer variation.

4. SignGraph: Efficient Pose-Based Graph Convolution [4]

Table 1: Summary of the Four Reviewed Papers

Ref.	Year / Venue	Input and Features	Method	Dataset(s)	Key Results	Limitations
[1] Tao et al.	2024, IEEE Access	Gloves, RGB, depth, skeleton (survey)	Traditional (HMM, CRF, DTW) and deep (CNN-RNN, 3D CNN, Transformer, GAN, GCN)	RWTH-PHOENIX, CSL, WLASL, AUTSL and others	Transformers reduce WER below 25 on continuous data; WLASL isolated accuracy only about 57%	Small datasets, no signer-independent tests, heavy models, no online recognition
[2] Miah et al.	2024, IEEE Access	67 body, face and hand key points plus joint motion	Two-stream GCN with attention and residual (GCAR), Sep-TCN	WLASL, PSL, MSL, ASLLVD	WLASL-100 top-1 63.25%; MSL 99.75%; ASLLVD 34.41%; 0.69 M parameters	Low accuracy for very large vocabularies; deep 22-stage design; offline use
[3] Kothadiya et al.	2022, IEEE Access	Static RGB image, 72×72 , 6×6 patches	Vision Transformer encoder with multi-head attention and MLP head	Static ISL, 36 classes	99.29% accuracy in 5 epochs	Static signs only; no temporal modelling

[4] Naz et al.	2023, IEEE Access	65 MediaPipe Holistic joints, 2D (x, y)	SignGraph: ResGCN with bottleneck blocks	WLASL-100/300/1000, LSA-64	Top-1 72.09 / 71.40 / 61.83%; LSA-64 100%; 20× fewer parameters than I3D	Isolated words only; ambiguous signs and lowresolution video cause errors
----------------	-------------------	---	--	----------------------------	--	---

IV. SIGN-TO-TEXT PIPELINE

The reviewed systems share the pipeline in Fig. 2. A camera captures frames, which are resized, flipped and converted to RGB. A pose estimator such as MediaPipe detects the hands and body and returns landmarks: MediaPipe Hands gives 21 points per hand and MediaPipe Holistic adds body and face points [4]. The landmark coordinates are normalised, for example shifted and scaled so that signer position and distance from the camera do not matter. For dynamic signs the motion between frames is added as a second feature, computed in [2] as $\text{Motion} = \text{text } X(t) - X(t+1)$ for each coordinate. system A feature model (CNN, GCN or Transformer) then learns spatial and temporal patterns, and a classifier outputs the sign. Isolated signs use a softmax classifier, while continuous signs need sequence methods such as Connectionist Temporal Classification (CTC) or an

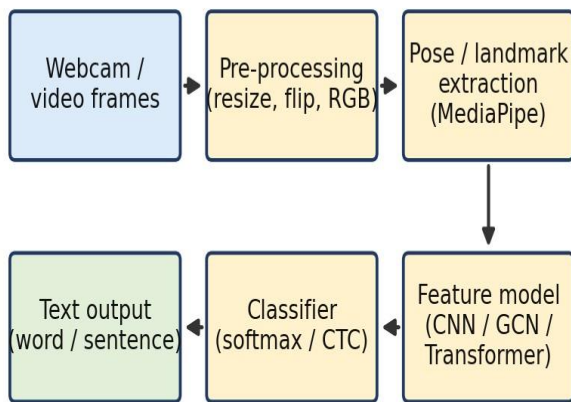


Fig.2. General pipeline of a vision-based sign-to-

Table II: Datasets Used in the Reviewed Papers

Dataset	Language	Type	Scale
WLASL [1]	American	Isolated	2,000 glosses, 21,083 videos, 119 signers
ASLLVD [2]	American	Isolated	2,745 signs, multiview video
LSA-64 [4]	Argentinian	Isolated	64 signs, 3,200 videos, 10 signers
AUTSL [1]	Turkish	Isolated	226 signs, 43 signers, 38,336 videos
PHOENIX-2014T [1]	German	Continuous	1,066 glosses, 8,257 videos, 9 signers
Static ISL [3]	Indian	Static images	36 classes, over 1,000 images per class

For isolated SLR the usual metric is top-k accuracy, where the correct sign must appear among the k highestscoring predictions. Continuous SLR uses Word Error Rate, $\text{WER} = 100 \times (\text{Substitutions} + \text{Deletions} + \text{Insertions}) / \text{Total words}$, where lower is better, and translation-style metrics BLEU and ROUGE [1]. A fair comparison should also state whether signers in the test set are different from those in training (signer independent), since the same signer in both sets makes results look better than real use [1].

V. COMPARATIVE ANALYSIS

Table III compares top-1 accuracy of RGB-based and pose-based methods on WLASL, as reported in [2] and [4], and Fig. 3 shows the WLASL-100 results. SignGraph is the best pose-based model on all three

subsets, and on WLASL-100 it also exceeds the RGB-based I3D model.

The advantage grows with vocabulary size: on WLASL300 SignGraph reaches 71.40% against 56.14% for I3D. The GCAR model improves the earlier Pose-TGCN on WLASL-100 and WLASL-300, but SignGraph is stronger because it uses a simpler 2D hand-and-body input with bottleneck blocks and residual connections.

Table 3: Top-1 Accuracy (%) on WLASL Subsets

Method	W-100	W-300	W-1000
Pose-GRU [4]	46.51	33.68	30.01
Pose-TGCN [4]	55.43	38.32	34.86
I3D, RGB [4]	65.89	56.14	47.33
GCAR two-stream [2]	63.25	43.80	–
SignGraph [4]	72.09	71.40	61.83

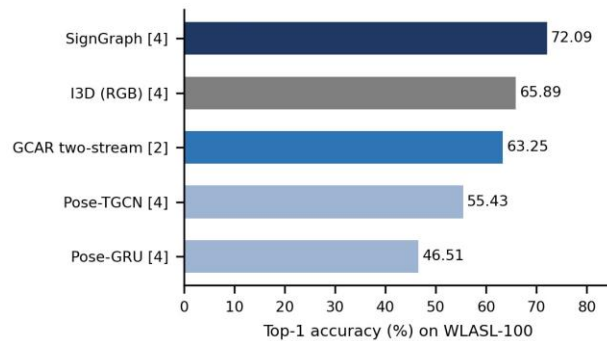


Fig.3. Top-1 accuracy of selected methods on WLASL-100 (values from [2], [4])

Computational cost differs even more. SignGraph needs 0.62 million parameters, 1.95 GFLOPs and 0.040 s per video against 12.4 million, 5.22 GFLOPs and 0.50 s for I3D [4], and GCAR has 0.69 million parameters [2]. In a small-data test on LSA-64, SignGraph reached 74.99% accuracy with only 10% of the training data while I3D reached 45.47% [4]. Pose input therefore suits real-time and low-resource use. The Vision Transformer in [3] reaches 99.29% but solves a different, easier problem, since it classifies 36 static signs from single images, so its accuracy cannot be compared directly with WLASL results. It shows, however, that attention based

models learn quickly even with few epochs. Taken together, the papers suggest that pose-based graph models are the best match for a webcam sign-to-text system, while Transformers become important when continuous sentences must be translated [1].

Challenges and Research Gaps

Large vocabulary and ambiguous signs: Accuracy falls sharply as the number of signs grows, for example from 63.25% on WLASL-100 to 24.10% on WLASL-2000 for [2] and 34.41% on ASLLVD. In [4] the main error source is different words that are signed almost identically.

Signer independence and unseen sentences: Most studies do not test signers or sentences absent from training, so real-world performance is uncertain [1].

Non-manual features: Facial expression, mouthing and body posture carry meaning, but hand-only systems ignore them. Adding face and body points helps [2] but increases noise and model size.

Real-time and lightweight deployment: Many accurate models are too complex for phones or webcams, and bidirectional models cannot run online because future frames are unavailable [1].

Continuous signing and grammar: Sign grammar differs from spoken grammar, so word-by-word output does not give natural sentences. Only a few reviewed systems handle continuous sentences [1].

Indian Sign Language data: Public ISL datasets are small and mostly static [3], so there is little data for dynamic, regional and sentence-level recognition.

Future Scope

Future systems can combine MediaPipe landmarks with lightweight graph or Transformer models so that sign-to-text runs in real time on a laptop or phone. Continuous recognition with CTC or encoder-decoder models, together with face and body cues, would move from single words to sentences. Larger, signer-independent datasets for Indian Sign Language, with several signers, backgrounds and cameras, are needed. Few-shot learning could add new signs without full retraining [1], and text-to-

speech output would complete a two-way communication aid. [5] World Health Organization, "Deafness and hearing loss," fact sheet. [Online]. Available:

VI. CONCLUSION

This paper reviewed four recent papers on vision-based sign language recognition. The survey [1] shows the move from gloves and hand-crafted features to deep learning and Transformers, and lists the open problems. The pose-based graph models [2], [4] reach high accuracy with very small models, and SignGraph outperforms RGB-based I3D on WLASL while being about twenty times smaller. The Vision Transformer [3] gives very high accuracy on static Indian signs. For a practical webcam-based sign-to-text system, landmark extraction with MediaPipe followed by a lightweight pose-based model is the most promising direction, and vocabulary size, signer independence and Indian Sign Language data remain the main challenges.

Acknowledgment

The author thanks the guide and the Department of Computer Engineering, Jaihind College of Engineering, Kuran, for their guidance and support.

REFERENCES

1. T. Tao, Y. Zhao, T. Liu and J. Zhu, "Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3398806.
2. A. S. M. Miah, M. A. M. Hasan, S. Nishimura and J. Shin, "Sign language recognition using graph and general deep neural network based on large scale dataset," IEEE Access, vol. 12, pp. 34553–34569, 2024.
3. D. R. Kothadiya, C. M. Bhatt, T. Saba and A. Rehman, "SIGNFORMER: DeepVision transformer for sign language recognition," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3231130.
4. N. Naz, H. Sajid, S. Ali, O. Hasan and M. K. Ehsan, "SIGNGRAPH: An efficient and accurate pose-based graph convolution approach toward sign language recognition," IEEE Access, vol. 11, pp. 19135–19147, 2023.