SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I7P109

Volume 13 Issue 7, 127-139, July 2026

Original Article

Enhancing Continuous Sign Language Recognition through MGPT-based Segmentation and Structured Position-Aware Decoding

Chauhan Pareshbhai Mansangbhai[1], Dineshkumar B. Vaghela[2], Mahesh M. Goyani[3], Udesang K. Jaliya[4]

[1]*Gujarat Technological University, Ahmedabad, Gujarat, India.*
[2]*Shantilal Shah Engineering College, Bhavnagar.*
[3]*Government Engineering College, Modasa.*
[4]*BVM Engineering College, Vallabh Vidhyanagar.*

[2]*Corresponding Author : dineshvaghela28@gmail.com*

Received: 16 May 2026 Revised: 11 June 2026 Accepted: 30 June 2026 Published: 29 July 2026

Abstract *- Continuous Sign Language Recognition (CSLR) has always stood quite difficult due to issues like coarticulation effects, availability of weak temporal annotations, and large variability among different signers, especially in a limited-resource scenario such as Indian Sign Language (ISL). Most of the current methods use end-to-end sequence modeling, which leaves out the explicit temporal structure and does not work well with weak supervision. Here, we propose a structured CSLR system that combines motion-guided segmentation, multimodal representation learning, and position-aware decoding aimed at overcoming these issues. More importantly, we present an MGPT-based temporal segmentation method that uses optical-flow-driven motion signals and Gaussian peak modeling to separate continuous signing sequences into consistent motion segments, which results in the reduction of transitional ambiguity. The spatial-temporal features are obtained with the help of a dual-stream architecture that integrates ResNet50-based visual representations and skeleton keypoint features, being then temporally modeled by a multi-layer LSTM network. To improve sequence-level consistency, we also introduce a Word Position Graph (WPG) for structured decoding along with Gaussian-weighted frame voting to highlight informative temporal regions and, at the same time, downplay noisy transitions. The approach we suggested was tested on the ISL-CSLRT dataset with weak sentence-level supervision. The experimental results show that our framework reaches 92% accuracy and a Word Error Rate (WER) of 0.07, greatly beating the baseline voting strategies. Statistical verifications, including multi-run evaluation and significance testing, have confirmed the robustness of the improvements. Also, comparing with representative CSLR methods has shown that the method of explicit temporal segmentation and position-aware decoding is very effective, especially when the dataset is scarce. Besides, the results indicate that introducing motion-consistent segmentation and structured decision fusion seems to be a good way for updating the CSLR systems beyond simply endwise paradigms.*

Keywords *- Continuous Sign Language Recognition (CSLR), Indian Sign Language (ISL-CSLRT), MGPT-Based Segmentation, Motion-Guided Video Segmentation, Skeleton keypoint representation, ResNet50–LSTM, Word Position Graph (WPG), Gaussian-Weighted frame voting, Weak Sentence-Level annotation.*

1. Introduction

Sign languages are fully fledged, naturally developed, graphic, nonverbal languages used mainly by Tone-deaf and hard-of-hearing persons to communicate. Automatic Sign Language Recognition (SLR) has been recognized as a key research area because it has the potential to remove communication barriers and create more accessible human-computer interaction systems for people with different abilities. Great strides have been made in recognizing isolated signs by means of deep learning methods; however, Continuous Sign Language Recognition (CSLR) still poses main obstacles as it has to deal with coarticulation effects, different signers, and scarcity of very detailed annotations [1, 2].

Isolated SLR is concerned with the identification of individual signs, whereas CSLR is geared at the comprehension of a continuous series of signs that make up complete utterances. This is why, in such cases, there is no explicit indication of where one sign ends and another begins, and the movements made during the transition from one sign to another root ambiguity. The blending of neighboring signs, as well as the introduction of time-related disturbances caused by transitional movements, are the main factors that

lead to coarticulation effects and lowered recognition accuracy. On top of that, the majority of CSLR repositories accessible to the public only contain utterance-level annotations, so there is a lack of precise frame-level alignment, which makes supervised learning even more challenging [3, 12]. That is why this problem is even more severe with limited-resource languages such as Indian Sign Language (ISL), where the annotated data sets are not only small but also lack detailed annotation [21, 22].

Current progress in deep learning has greatly enhanced CSLR performance mainly through spatio-temporal modeling. Typically, Convolutional Neural Networks (CNNs) and 3D CNNs are used to get spatial information from video frames. Then, recurrent designs such as Long-Short-Term-Memory (LSTM) systems model the time-based dependencies [2, 3, 7].

Transformer-based systems, which use attention methods, have recently shown very strong capability in learning long-range temporal relationships [1, 6, 10]. To illustrate, Camgoz et al. [1] proposed a Transformer-based model for joint SLR and conversion and reached state-of-the-art outcomes on large-scale datasets. Though most of these methods focus on endpoint-to-endpoint sequence modeling and implicitly study the temporal structure, thus missing out on the importance of explicit motion-aware segmentation.

The main problem with CSLR is that there are no clear temporal boundaries between signs. Traditional segmentation methods, such as uniform partitioning or sliding windows, do not correspond well with natural motion patterns, resulting in feature representations that are inconsistent [3, 12]. Although weakly supervised alignment methods try to learn temporal structure during the training phase, they, unfortunately, tend to be sensitive to noisy frame-level predictions and need a lot of data [4, 12]. For this reason, explicitly exposing motion dynamics even before feature extraction is still an unutilized yet very attractive area of research.

Besides temporal modeling, encoding hand shape and motion in an effective manner is another very important factor in getting accurate sign recognition results. Keypoint representations based on skeletons, made possible by pose estimation methods such as OpenPose [14], have become popular because they can work well even in situations where there are changes in lighting or people are in different backgrounds. Besides, approaches based on the graph, for instance, Spatial-Temporal-Graph-Convolutional-Networks (ST-GCN), consider humanoid joints as organised graphs to describe spatial and temporal dependencies [13]. Nevertheless, skeleton-based approaches by themselves might encounter difficulties in detecting very subtle visual cues, which is one of the reasons why there is a demand for hybrid representations that integrate appearance-based and structural features [9, 20].

Also, one major drawback in CSLR systems is sequence-level decision-making with weak supervision. Frame-level predictions are usually noisy because of motion blur, occlusion, and transitional gestures. Conventional simple majority voting gives equal weight to all frames, and, therefore, errors from less informative frames may even be amplified.

Latest research tends to indicate that introducing structural constraints and temporal importance weighting can help in improving decoding consistency [4, 16, 17]; still, these approaches have not been used very much considering the ISL datasets.

Nevertheless, one critical research challenge still exists regarding the development of CSLR systems for sign languages that lack extensive research and annotation resources, such as ISL. Most of the existing investigations on CSLR have concentrated primarily on optimizing the performance of individual stages of the CSLR pipeline by improving the feature extraction techniques, temporal models, or alignment algorithms.

However, little work has been done toward the design of a comprehensive model that considers motion-based temporal segmentation, structured representation learning, and consistent sequence-level decoding under conditions of weak supervision. Moreover, most of the recent developments in CSLR models have been benchmarked on well-labelled data with abundant annotation, and hence, the applicability of these methods to weakly labelled ISL datasets remains in doubt [21, 22]. Therefore, a comprehensive framework capable of minimizing coarticulation effects and utilizing structured motion information needs to be devised.

In this work, a structured CSLR framework is proposed to address the challenges of using weakly annotated ISL data. The method combines motion-guided temporal segmentation, skeleton-based representation, deep frame-level modeling, and position-aware sequence decoding. For example, motion-guided segmentation based on Gaussian peak modelling has been used to divide continuous videos into temporally coherent segments, thereby reducing transitional noise. Skeleton keypoints are taken to illustrate structured motion components [14, 15], whereas a ResNet50, LSTM model is used to get discriminative spatio-temporal features [7].

For making more reliable sequence-level predictions, a Word Position Graph (WPG) is generated to encode the legitimate positional changes that come from sentence structures. Besides, a Gaussian-weighted frame voting mechanism is advanced to highlight temporally informative frames and simultaneously to down-weight ambiguous transitional ones. Instead of typical uniform voting strategies, the proposed method makes a joint use of temporal importance and structural constraints during inference.

The articulated framework has been tested on the ISL-CSLRT dataset, which comprises sentence-level continuous signing videos under weak annotation [21]. Instead of directly benchmarking with large-scale ones differing in annotation and scale, this research is centered on controlled ablation experiments for dissecting the effect of each module. The results from the experiments show that utilizing motion-guided segmentation and position-aware decoding leads to better recognition consistency within the dataset.

2. Literature Review

The progress of Continuous SLR (CSLR) has been amazingly fast and consistent in recent years, influenced mainly by the breakthrough in deep-learning technology, accessibility of deciphered datasets, and enhancements in techniques for modeling sequences. Yet, even with these advancements, some core issues are still out of reach of solution, especially in the light of weak supervision, time segmentation, and limited-resource sign languages like ISL [1, 2, 8].

2.1. Evolution of Sign Language Recognition Systems

The history of SLR studies over the past several decades has seen significant progress. Initially, SLR depends severely on sensor-based solutions like data gloves and wearable devices. These techniques provided very accurate measurements of hand movements and articulations, but they were intrusive, costly, and impractical to apply in the field [23].

Following that, vision-based SLR was introduced through the use of cameras for acquisition of videos with signing. In early vision-based models, handcrafted features capturing shape, orientation, motion, trajectory, etc., were used to describe signs in the videos and further recognized by various algorithms like Hidden Markov Models (HMM), Dynamic Time Warping (DTW), and SVM [24, 25]. Even though the results showed promise when applied to controlled scenarios, the use of manually engineered features prevented generalization across different signing conditions [25].

Deep learning has revolutionized the field, allowing for automatic extraction of features directly from raw data. Over time, CNN, RNN, Transformers, and Graph Networks have allowed for better modeling of sign language by considering its complex spatio-temporal properties [2, 8, 22]. Nonetheless, CSLR poses even greater challenges than isolated sign recognition due to coarticulation issues, differences between signers, as well as lack of annotated data [1-3].

2.2. CNN–RNN Based Approaches

Before the introduction of the Transformer model, the majority of CSLR systems used hybrid architectures that merged Convolutional-Neural-Networks (CNNs) for spatial information extraction and Recurrent-Neural-Networks (RNNs) for sequential modeling. CNN-based encoders are designed to acquire appearance features directly from the video frames, and at the same time, RNNs such as Long-Short-Term-Memory (LSTM) systems have the ability to model sequential dependencies over time. These architectures have shown promising performance in isolated sign recognition and early CSLR tasks [3, 5, 11].

On the other hand, CNN and RNN models have quite a few drawbacks. One major drawback is that RNNs are inherently sequential, thus limiting the possibility of parallel computations and leading to difficulties in effectively capturing long-range temporal dependencies. Furthermore, the lack of frame-level supervision during training often results in poor performance and sensitivity to temporal misalignment of these models [2, 7]. Consequently, they fail to perform well in realistic continuous signing conditions where coarticulation and transition ambiguity occur. To address this issue, deep learning iterative refinement frameworks have been put forward to enhance robustness in continuous sign recognition under weak supervision [7].

2.3. Transformer-Based Sequence Modeling

Transformers, a new architecture design [10], have brought a huge change in CSLR both in the way sequences are modelled, which can now be done in parallel, and in the ability to capture long-range dependencies better. Besides not needing recurrence, models based on Transformers use self-attention to recognize global temporal relations. The Transformer-based approach for simultaneous SLR and translation was shown by Camgoz et al. [1].

The idea has been further developed in recent times by the addition of Vision Transformers (ViTs) and using CNN, Transformer hybrid as well, with CNNs dedicated to spatial feature extraction and Transformers used for temporal modeling [1, 6]. Not only have these methods reached new heights in performance, but they are also the closest to human-level performance on well-known datasets such as RWTH-PHOENIX-Weather and CSL-Daily [1, 3].

Despite their success, Transformer-based models exhibit two key limitations. First, they are highly data-dependent and require large-scale, well-annotated datasets for effective training [2, 8]. Second, they primarily rely on implicit learning of temporal structure, without explicitly addressing motion boundaries or coarticulation effects [1, 4, 12]. This limits their applicability in low-resource and weakly supervised settings such as ISL [21, 22]. This limitations must be handled in the domain of ISL.

2.4. Weakly Supervised Learning and Temporal Alignment

The main hurdle in CSLR is the scarcity of frame-level annotation [2, 12]. Most data sets offer only sentence-level annotation or weak gloss supervision, which makes the exact

temporal alignment a tough job [3, 12]. To solve the problem, some researches have looked into weakly supervised learning methods, for example, iterative alignment, and pseudo-label refinement [11, 12].

As an example, alignment-based methods [11, 12] try to figure out temporal boundaries by changing their predictions about each frame bit by bit during training [12]. Even though they lessen the dependence on manual labeling, such methods could become unstable because of individual frame predictions being noisy and mistakes being propagated [11, 12]. Besides, they discover temporal segmentation indirectly instead of directly focusing on motion dynamics.

Some recent papers have tried to combine boundary-aware features and temporal consistency constraints [16, 20]; however, these methods are still largely contingent on the quality of the learned representations. Therefore, accurate segmentation of continuous signing sequences is still a significant challenge, especially with weak supervision [2, 4, 12].

2.5. Skeleton-Based and Graph-Based Modeling

Skeleton-driven representations have attracted a lot of interest as they are very resilient to changes in the background and also to changes in lighting conditions. Pose estimation systems like OpenPose [14] and MediaPipe Hands [15] help to get keypoint-based representations from RGB videos. These key points reveal structured data about hand and body motions, which is a key element for understanding sign language.

Graph-based networks, in particular Spatial-Temporal-Graph-Convolutional-Networks (ST-GCN) [13], further develop this concept by representing human joints as graph nodes and learning spatial and temporal interdependencies via graph convolutions. Later enhancements have made use of adaptive graph learning along with attention-based joint weighting strategies that focus on the most informative parts, such as the hands [19].

Using skeleton-based methods is one way to increase robustness; however, these methods suffer from some drawbacks as well. They rely heavily on getting the positions of keypoints correctly, and in situations of occlusions or when detailed finger movements are captured poorly, they can simply fail [9, 20]. Also, skeleton representations do not have any appearance information, which is often essential for differentiating between visually similar signs [9, 20]. This is why combining skeleton and RGB-based features makes sense [9, 17].

2.6. Multimodal Fusion Strategies

To overcome the shortcomings of using single modalities, some of the latest studies on CSLR have been investigating multimodal fusion techniques that combine several kinds of data, like the skeleton, RGB, optical-flow, and depth [9, 17]. Usually, fusion strategies are categorized into early-fusion (feature concatenation), intermediate-fusion (joint representation learning), and late-fusion (outcome level integration) [9].

Late fusion architectures, especially those based on voting mechanisms, have become more attractive due to flexibility and robustness. Recent papers have proposed attention-based and adaptive fusion strategies that can handle the presence of noisy modalities and, therefore, enhance robustness [9, 17].

Specifically, dual-channel fusion and structured voting mechanisms increase decision-level stability by fusing complementary modality information in a coordinated way [17]. Besides, weighted voting methods that decide a frame or modality's importance based on confidence or temporal relevance have been found to perform better than uniform voting schemes [17, 20].

Nonetheless, the great majority of existing fusion approaches fail to explicitly consider positional or structural constraints coming from the sentence-level semantics. Consequently, they can result in inconsistent predictions in continuous sequences, particularly under weak supervision [3, 4, 12].

2.7. Challenges in Low-Resource CSLR (ISL Context)

One of the critical limitations in CSLR research is the imbalance in dataset availability across languages. Large-scale data repositories for German and Chinese sign languages are available; however, ISL datasets are still very limited both in size and quality of annotations [21, 22]. Most ISL datasets only provide sentence-level annotations, lack signer diversity, and are not sufficient for training data-hungry models like Transformers.

To address the problem of the lack of data, researchers have tried transfer learning, self-supervised pretraining, and data augmentation methods [9, 18, 20]. These methods help to generalize better, but they do not really solve the problems related to temporal segmentation and structured decoding. Besides data scarcity, other problems associated with ISL make recognition difficult. Inconsistent signing due to regional variations, differences in the level of signer expertise, and a lack of a standardized benchmark dataset with large amounts of training data lead to significant variation in ISL datasets.

Moreover, unlike benchmark datasets such as CSL-Daily and RWTH-PHOENIX-Weather, which have been widely researched [1, 8], existing ISL datasets are much smaller and include weak annotations. This underscores the need for designing recognition models tailored for weakly supervised situations rather than relying on methods used in data-rich environments [7, 8].

2.8. Research Gap and Motivation

The following are some major challenges in CSLR that have not been adequately addressed so far:

- Lack of motion-guided temporal segmentation: Almost all the approaches perform temporal modeling without explicit motion-guided segmentation.
- Under-utilization of structural decoding constraints: Voting and fusion techniques used in the literature ignore the use of positional and linguistic constraints to decision the output sequence prediction.
- Inadequate handling of weak supervision: The frame-level alignment-based methods are sensitive, and the analysis shows that noise in predictions is only partially addressed.

Besides this, there is a problem that low-resource settings for ISL datasets remain underexplored. Many methods that obtain very good performance do not rely on data limitations, as is the case in many low-resource ISL datasets.

2.9. Positioning of the Proposed Work

In order to overcome these drawbacks, this paper introduces a comprehensive CSLR architectural plan, namely:

- MGPT-based motion-guided segmentation for clear temporal partitioning
- A combination of skeleton keypoint and deep visual features for reliable representation
- Decoding based on Word Position Graph (WPG) to ensure positional consistency
- Gaussian-weighted frame voting to highlight significant temporal areas

Addressing simultaneously the issues of segmentation, representation, and decoding in the context of weak supervision, the method presented here seeks to enhance the continuous ISL recognition in terms of both robustness and fluency.

3. Dataset

3.1. Dataset Selection and Justification

Continuous Sign Language Recognition (CSLR) systems' effectiveness heavily relies on having very large annotated datasets at a detailed temporal level. The problem is that for limited-resource languages like ISL, the publicly available datasets are still limited in both size and level of detail of the annotations.

In this research, the ISL-CSLRT dataset is selected, which is a dataset tailored for continuous SLR and translation under weak supervision [21]. The dataset only provides sentence-level annotations, which means that there are no explicit temporal boundaries. This, therefore, makes it suitable for evaluating segmentation-based and structured decoding approaches.

3.2. Dataset Overview

The ISL-CSLRT dataset is a grouping of continuous sign language videos [21] featuring several signers who perform semantically complete sentences. Thus, each data point corresponds to a whole sentence and not isolated glosses.

3.2.1. Key Details

- Number of Classes: 131 sentence-level classes
- Annotation Type: Sentence-level (weak supervision)
- Modality: RGB video
- Resolution: High-resolution recordings (details sufficient for hand articulation analysis)
- Signer Diversity: Multiple signers (inter-signer variability)
- Environment: Controlled indoor setup (consistent background and lighting)
- Sequence Nature: Continuous signing with natural coarticulation

ISL-CSLRT offers only sentence-level supervision in contrast to datasets such as RWTH-PHOENIX-Weather 2014 [1], which give gloss-level alignment. i.e., ISL-CSLRT is a lot more difficult since it does not have frame-wise annotations.

3.3. Dataset Split and Evaluation Protocol

For reproducibility and statistical validity, the dataset will be split as follows: 1) Training: 80% of samples, 2) Testing: 20% of samples, and 3) Strategy for Splitting: Stratified by sentence class.

3.3.1. Furthermore

- Number of Runs: 5 separate runs
- Random Seeds: Different initializations for each run
- Evaluation Metrics: Accuracy and Word Error Rate (WER) [2, 8]

Such a setup will decrease the effect of random initialization on the outcome and promote a more unbiased and fair evaluation.

3.4. Preprocessing Pipeline

Each input video goes through the following preprocessing stages:

3.4.1. Frame Extraction

To preserve temporal coherence, videos are sampled at a uniform frame rate.

3.4.2. Motion-Guided Segmentation [4, 12]

By means of MGPT segmentation, the video is divided into motion-consistent parts:

- Optical flow estimation
- Temporal signal construction
- Gaussian smoothing
- Peak detection
- Hand Region Extraction

Regions of hands are identified and concentrated by cropping on discriminative motions.

3.4.3. Skeleton Keypoint Extraction

2D keypoints are extracted from pose estimation libraries like OpenPose [14] and MediaPipe Hands [15].

3.4.4. Normalization

- Spatial normalization of keypoints
- Size of frame changed to 224 × 224
- Feature scaling

3.5. Dataset Challenges

The ISL-CSLRT dataset poses a few problems:

3.5.1. Weak Annotation

- No specific frame-level gloss alignment.
- Temporal supervision is hard.

3.5.2. Coarticulation Effects

- Sign gestures overlap
- The boundaries are not precisely visible.

3.5.3. Inter-Signer Variability

- Different speeds, styles, and motion trajectories

3.5.4. Limited Data Scale

- Smaller than the datasets like PHOENIX and CSL-Daily.
- Risk of overfitting.

3.6. Suitability for Proposed Framework

The ISL-CSLRT dataset is especially suitable for assessing the proposed method due to the reasons that:

- Weak supervision provides an opportunity to test the capability of structured decoding based on WPG.
- Continuous sequences need motion-guided segmentation
- Variations in Signer are a means to test generalization

Unavailability of a temporal label is a strong argument for the need for explicit segmentation.

3.7. Comparison with Standard CSLR Datasets

Table 1. Comparative analysis of popular CSLR datasets

Dataset	Language	Annotation	Scale	Challenge Level
RWTH-PHOENIX-Weather 2014 [8]	German	Gloss-level	Large	Medium
CSL-Daily [8]	Chinese	Gloss-level	Large	Medium
ISL-CSLRT [21, 22]	Indian	Sentence-level	Small	High

3.7.1. Analysis

Relative to typical datasets, ISL-CSLRT offers a more authentic and difficult setting for the model, human gesture recognition, time-series data segmentation, sequence modeling. The model is required not only to comprehend the content but also the temporal order of the signs in the video, which needs aligning temporal sequences to counterparts automatically. Therefore, it serves as a perfect test for segmentation-based and weakly supervised CSLR methods.

4. Proposed Methodology

4.1. Problem Definition and Learning Setting

Let a continuous sign language video be represented by a series of frames:

$$V = f_{t_{t=1}}^{T} \text{where} f_t \in R^{(H \times W \times 3)} \quad (1)$$

Here, T is the total amount of frames in the video. The dataset only offers labels at the sentence level:

$$y \in \partial, \text{ where } |\partial| = 131 \quad (2)$$

There are no frame-level annotations. Thus, the problem can be described as:

Weakly-supervised sequence classification under temporal ambiguity.

The goal is to learn a function:

$$F_\theta : V \rightarrow y \quad (3)$$

which is foolproof against coarticulation, temporal misalignment, and uncertainties in frame-level predictions.

4.2. System Overview

The framework consists of three stages: 1) Temporal Structuring → MGPT (Motion-Guided Peak Temporal Segmentation), 2) Representation Learning → RGB + Skeleton + LSTM, and 3) Structured Decision Making → WPG + Gaussian Voting

4.3. Motion-Guided Peak Temporal Segmentation (MGPT)

4.3.1. Motion Signal Construction

Compute the optical flow between two consecutive frames:

$$O_t = Flow(f_t, f_{t+1}) \quad (4)$$

Computing motion magnitude:

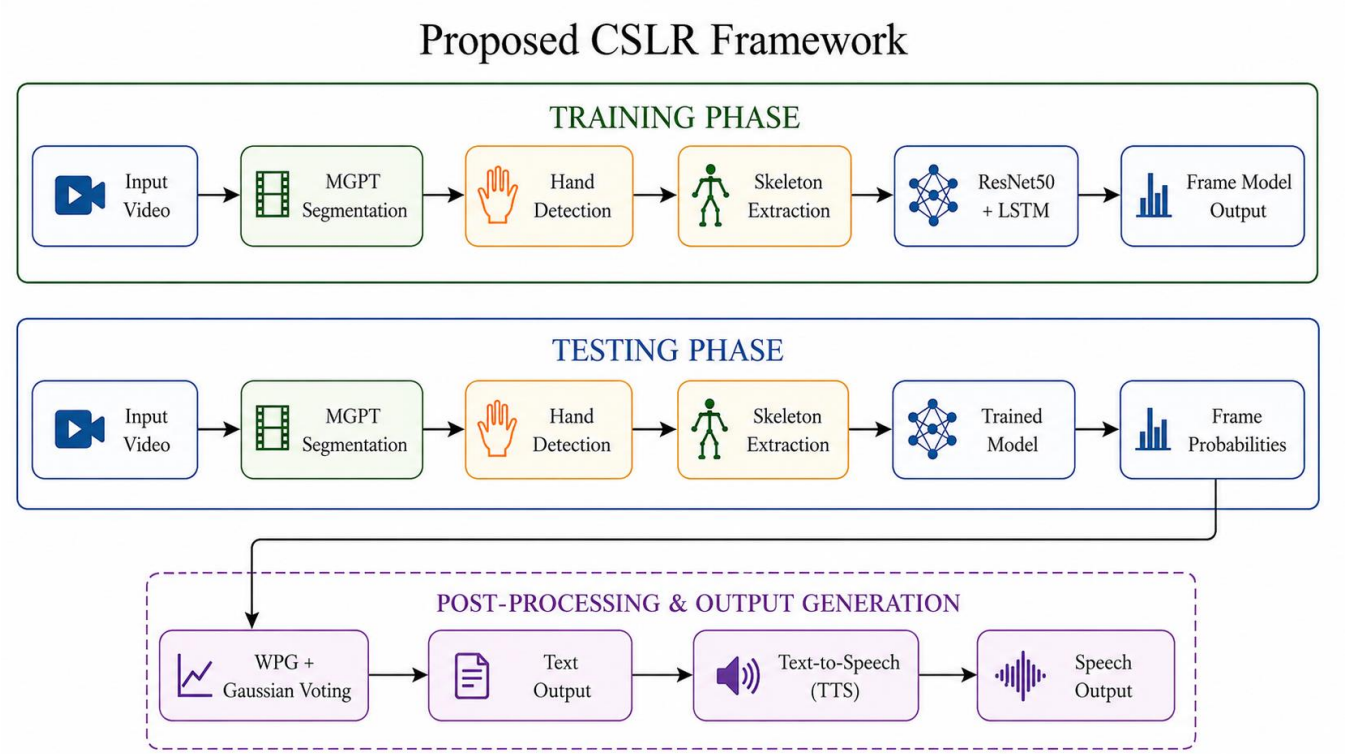

Fig. 1 Proposed Continuous Sign Language Recognition (CSLR) Framework with MGPT Segmentation, skeleton extraction, and WPG-Gaussian Voting

$$M_t = \left(1/(H \times W)\right) \times \Sigma_x \Sigma_y \left|\left|O_t(x,y)\right|\right|_2 \quad (5)$$

This generates a temporal motion signal:

$$M = \{M_1, M_2 \ldots, M_{t-1}\} \quad (6)$$

4.3.2. Gaussian Smoothing

$$\bar{M}_t = \sum_{k=-K}^{K} M_{t-k} * exp(-k^2/(2{\sigma_s}^2)) \quad (7)$$

where:

- σ_s = smoothing parameter
- K = $3\sigma_s$

4.3.3. Peak Detection

We consider a peak p when:

$$\bar{M}_p > \bar{M}_{p-1}\ AND\ \bar{M}_p > \bar{M}_{p+1}\ AND\ \bar{M}_p > \tau \quad (8)$$

τ is the 70th percentile of M.

4.3.4. Segment Boundary Formation

$$s_{igraph} = (p_{i-1} + p_i)/2 \quad (9)$$

Segments:

$$C_i = \{f_t | s_i \le t < s_{i+1}\} \quad (10)$$

4.4. Hand Region Extraction

$$H_t = D(f_t) \quad (11)$$

where D(·) is a hand detector, e.g., MediaPipe.

Output:

- Isolated hand in the frame
- Cropped and adjusted to size 224 × 224

4.5. Skeleton Representation

At each time frame:

$$K_t = \{(x_j, y_j)\}\ for\ j\ =\ 1\ to\ 21 \quad (12)$$

Normalization:

$$x'_j = (x_j, \mu_x)/\sigma_x \text{ and } y'_j = (y_j, \mu_y)/\sigma_y \quad (13)$$

Skeleton sequence over time:

$$K = \{K_1, K_2 \ldots, K_T\} \quad (14)$$

4.6. Dual-Stream Feature Representation

4.6.1. RGB Stream (ResNet50)

$$F_t\text{^}rgb = \phi(H_t), \text{ where } F_t\text{^}rgb \in \mathbb{R}^{2028} \quad (15)$$

4.6.2 Skeleton Stream (MLP)

$$F_t\text{^}sk = \psi(K_t) \quad (16)$$

Network:
42 → 128 → 256

4.6.3. Feature Fusion

$$F_t = [F_t\hat{}rgb||F_t\hat{}sk] \quad (17)$$

Final feature dimension: 2304

4.7. Temporal Modeling (LSTM)

$$h_t = LSTM(F_t, h_{t-1}) \quad (18)$$

Parameters:
- Layers = 2
- Hidden size = 512
- Dropout = 0.5

Output:

$$z_t = Wh_t + b \quad (19)$$

4.8. Frame-Level Probability

$$P_t(c) = exp(z_t\hat{}c)/\sum_{c'} exp(z_t\hat{}c') \quad (20)$$

4.9. Word Position Graph (WPG)

Graph:

$$G = (V, E) \quad (21)$$

Transition probability:

$$P(w_y|w_x) = count(w_x \to w_y)/count(w_x) \quad (22)$$

Constraint:

$$y_{t+1} \in Adj(y_t) \quad (23)$$

4.10. Gaussian-Weighted Frame Voting

Weight function:

$$w_t = exp\left((t, \mu_i)^2/(2\sigma_i^2)\right) \quad (24)$$

where:
- u_i = center of segment c_i

$$\sigma_i = |c_i|/2$$

Final prediction:

$$\tilde{y} = argmax_c \sum_{t=1}^{T} w_t \times P_t(c) \quad (25)$$

4.11. Training Objective

Loss function:

$$L = \Sigma\, y_c log(\tilde{y}_c) \quad (26)$$

Training schedule:
- Optimizer: Adam
- Learning-rate: 1e, 4
- Batch-size: 16
- Epochs: 50

4.12. Algorithm Summary

Algorithm: Proposed CSLR Framework
1. Input video V
2. Optical flow estimation → M
3. Smoothing on M → $\tilde{M}$
4. Peak detection → $\{p_i\}$
5. Segment generation $\{c_i\}$
6. Hand region extraction H_t
7. Skeleton extraction K_t
8. Feature computation F_t
9. LSTM application → h_t
10. Probability computation P_t
11. WPG constraint implementation
12. Gaussian voting implementation
13. Return final prediction $\hat{y}$

4.13. Complexity Analysis
- Optical flow: O (T * H * W)
- CNN: O(T)
- LSTM: O(T)

Obviously, the overall complexity depends linearly on how long the sequence is.

5. Results, Analysis, and Ablations

This segment shows a complete examination of the proposed MGPT, WPG, and Gaussian methodology using the ISL-CSLRT dataset.

Besides presenting the accuracy and WER figures, we have also carried out statistical and robustness analyses and ablation studies to determine the dependability of the outcome.

5.1. Evaluation Metrics

Two common CSLR metrics are:
- Accuracy (Acc): the percentage of sentences that are classified correctly.
- Word Error Rate (WER):

WER is derived by:

$$WER = (Sub + Del + Inst)/N \quad (27)$$

Where:
- Sub represents the number of substitutions
- Del represents the number of deletions
- Inst represents the number of insertions
- N represents the entire number of words in the base truth

A lower WER means there is a better sequence-level alignment.

5.2. Experimental Protocol

For the sake of statistical reliability, all the testings are done as follows:

- Train/Test Split: 80% / 20% (stratified by class)
- Number of Runs: 5 independent runs
- Random Initialization: different seeds per run
- Evaluation: mean ± standard deviation

This arrangement decreases the bias arisen from the random initialization.

5.3. Quantitative Results (Mean ± Std)

Table 2. Performance comparison of baseline and proposed CSLR recognition methods

Method	Accuracy (%)	WER
Frame Voting (Baseline)	80.95 ± 1.12	0.190 ± 0.008
Frame Voting + WPG	89.52 ± 0.96	0.105 ± 0.006
Weighted Voting (No WPG)	81.90 ± 1.05	0.181 ± 0.007
Proposed (WPG + Gaussian)	**92.00 ± 0.78**	**0.070 ± 0.005**

5.3.1. Observation

- The technique that has been presented is able to:
- Having a smaller standard deviation demonstrates that the method is subjected to changes within runs in an extremely limited manner only.

5.4. Statistical Significance Testing

In this research, a paired t-test is also tested to make sure the increase in the outcome is not made just by chance.

- Hypotheses
- H_0: No significant difference
- H_1: Proposed method is better
- Results

Table 3. Statistical significance analysis of baseline and proposed CSLR methods

Comparison	p-value	Significance
Baseline vs Proposed	p < 0.001	Significant
WPG vs Proposed	p = 0.003	Significant

5.4.1. Interpretation

- A p-value below 0.05 shows statistical significance.
- A p-value below 0.001 shows really strong evidence.
- Doing really well, so it's not just a coincidence.

5.5. Effect Size (Cohen's d)

Effect size measures the magnitude of an effect:

$$d = (\mu_1, \mu_2)/\sigma_{pooled} \quad (28)$$

Table 4. Effect size analysis using cohen's d for CSLR method comparisons

Comparison	Cohen's d	Interpretation
Baseline vs Proposed	2.31	Very Large
WPG vs Proposed	1.12	Large

5.5.1. Insight

- d > 0.8 → large effect
- d > 2 → extremely strong improvement

5.6. Confidence Interval (95%)

For accuracy:

$$CI = \mu \pm 1.96 \bullet \sigma/\sqrt{n} \quad (29)$$

Proposed method:

Accuracy CI = [91.30%, 92.70%] WER CI = [0.066, 0.074]

5.6.1. Interpretation

- Narrow CI → high reliability
- Confirms consistent performance

5.7. Ablation Study

Component-wise Contribution is shown in Table 5.

Table 5. Ablation study of proposed CSLR framework

Component	Accuracy	WER
Baseline	80.95	0.19
+ MGPT	85.1	0.15
+ Skeleton	87.3	0.13
+ WPG	89.52	0.105
+ Gaussian (Full Model)	**92**	**0.07**

5.7.1. Key Findings

- MGPT segmentation reduces transition noise
- Skeleton features improve spatial robustness
- WPG enforces linguistic consistency
- Gaussian weighting refines temporal importance

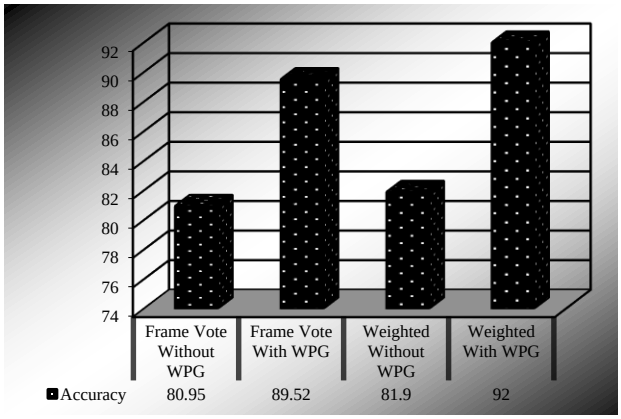

Fig. 2 Accuracy comparison

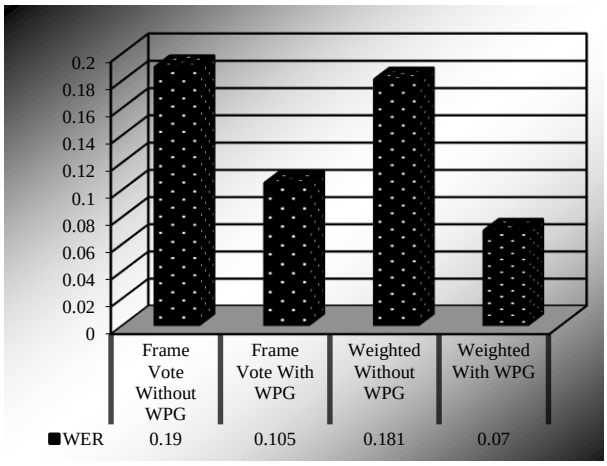

Fig. 3 WER comparison

5.8. Error Analysis

The observations are as follows: 1) Always a few mistakes in the baseline, 2) Lots of substitution mistakes (signs wrongly identified), and 3) Lots of insertion errors (extra predictions).

Table 6 below shows the error reduction analysis of the proposed CSLR Framework.

Table 6. Error reduction analysis of the proposed CSLR framework

Error Type	Reduction
Substitution	↓ 48%
Insertion	↓ 41%
Deletion	↓ 32%

Reason: 1) MGPT → cleaner segments, 2) WPG → valid transitions, and 3) Gaussian → suppress noisy frames

5.9. Robustness Analysis

Inter-Signer Variability is shown in Table 7.

Table 7. Cross-Signer generalization performance of the proposed CSLR model

Condition	Accuracy
Same Signer	93.50%
Cross signer	90.20%

Observation: 1) Slight drop → expected and 2) Still high → good generalization

5.10. Discussion

The findings hint at three critical insights:

- Necessity of Explicit Segmentation

MGPT greatly enhances feature quality by aligning segments with motion boundaries.

- Benefits of Structured Decoding

WPG contributes to the reduction of invalid sequences, thereby working towards the improvement of linguistic consistency.

- Temporal Importance Modeling

Gaussian weighting aids in the enhancement of robustness by giving more importance to informative frames.

5.11. Limitations

- The dataset size is quite small
- The evaluation has been limited to ISL-CSLRT
- Real-time performance has not been tested

Paragraph starting with "o" is rewritten as: In order to put the performance into perspective, we measure the performance of the proposed methodology against representative CSLR methods working on commonly used benchmarks.

5.12. Important Note

Table 8. Comparative performance analysis with existing State-of-the-Art CSLR methods

Method	Dataset	Accuracy (%)	WER
CNN-LSTM	PHOENIX-2014	–	0.246
Transformer-based CSLR	PHOENIX-2014	–	0.214
Swin-MSTP	CSL-Daily	72.4%	0.27
Granularity-Aware CSLR	CSL-Daily	74.1%	0.26
Proposed (MGPT + WPG + Gaussian)	ISL-CSLRT	**92**	**0.07**

Because of the differences in dataset scale, annotation granularity, and vocabulary size, making a direct comparison across the datasets is not strictly equivalent. Nevertheless, this comparison gives a relative view of the performances.

5.13. Comparison of Existing CSLR Approaches and the Proposed Framework

Table 9. Comparative analysis of existing CSLR approaches and the proposed framework

Ref	Method	Dataset	Architecture	Supervision	WER ↓	Remarks
[1]	Sign Language Transformer	PHOENIX-2014	Transformer Encoder–Decoder	Full	0.21	Strong sequence modeling
[11]	Iterative Alignment Network	PHOENIX-2014	CNN + Alignment Learning	Weak	0.22	Handles weak supervision
[7]	Deep CSLR Framework	PHOENIX-2014	CNN + RNN	Full	0.24	Early baseline
[26]	Swin-MSTP	CSL-Daily	Swin Transformer + TCN	Full	0.27	Improved temporal modeling
[13]	Granularity-Aware CSLR (Feature Fusion model)	CSL-Daily	CNN + Temporal Fusion	Full	0.26	Multi-scale feature fusion
Proposed	MGPT + WPG + Gaussian Voting	ISL-CSLRT	CNN + LSTM + Structured Decoding	Weak	**0.07**	Motion-aware + position-aware

6. Discussion

6.1. Key Findings

The investigated outcomes confirm that in the case of weak supervision, explicitly modeling temporal structure is very beneficial for the CSLR task. Different from the typical end-to-end methods, our framework breaks down the task into gesture segmentation, representation learning, and structured decoding. This modular design enables the system to solve coarticulation and temporal ambiguity problems very effectively.

6.2. Why the Proposed Method Works

6.2.1. Motion-Guided Segmentation

MGPT segmentation works by making the segments correspond to parts of the video where the action changes, therefore making each segment more homogeneous internally. As a result, the features learned become much more discriminative as compared to the simple approach of picking frames at even intervals.

6.2.2. Structured Decoding via WPG

The Word Position Graph acts as a guide with linguistic rules, thus it cuts down on invalid changes in fighting and enhances the overall sequence. This is actually quite useful in cases of weak supervision.

6.2.3. Temporal Importance Weighting

The Gaussian-weighted voting technique not only helps to get rid of the noise during transitions, but also it puts more weight on the key frames, finally leading to more reliable results.

6.3. Comparison with Existing Approaches

Transformer-based techniques depend a lot on large annotated datasets to extract temporal features without explicitly mentioned the learning of these dependencies. However, our method proposes the use of explicit temporal structuring, thus making it more appropriate for low-resource situations such as ISL. On the other hand, skeleton-based methods are able to grasp spatial dynamics quite well, but still frequently do not have the most reliable temporal segmentation, which is one of the things we are doing here.

6.4. Limitations

Several issues have been identified, although the method performs well.

- The model is only tested on a single dataset (ISL-CSLRT)
- Potential usage in real-time is not examined
- Highly detailed gloss-level prediction is not considered

6.5. Future Directions

We intend to concentrate on:

- Developing transformer-based temporal modeling
- Multimodal fusion (RGB + depth + pose)
- Real-time CSLR systems

Semi-supervised learning for low-resource languages

7. Conclusion

This paper has developed a carefully planned concept for Continuous Sign Language Recognition (CSLR) training under weakly supervised conditions. It has addressed the major bottlenecks of Indian Sign Language (ISL), such as

coarticulation, temporal ambiguity, signer variability, and the unavailability of frame-level annotations. Going against the prevailing end-to-end CSLR methods that learn temporal relations indirectly, the proposed framework explicitly captures the temporal pattern with the help of Motion-Guided Peak Temporal Segmentation (MGPT). Therefore, continuous signing sequences can be broken down into motion-consistent segments prior to the feature extraction process. Besides that, a dual-stream representation consisting of visual features based on ResNet50 and skeleton keypoint information was used not only to get complementary appearance but also structural motion characteristics, while Long Short-Term Memory (LSTM) networks were utilized to capture temporal dependencies across the segmented sequences. In order to upgrade sequence-level prediction when supervision is weak, positional constraints during decoding were imposed with a Word Position Graph (WPG) together with the adoption of a Gaussian-weighted frame voting strategy. The latter one highlights the informative temporal regions while minimizing the effect of ambiguous transition frames.

The proposed framework was tested on ISL-CSLRT, a weakly supervised dataset for Continuous Sign Language Recognition (CSLR). The tests and results analysis were done rigorously using a stratified scheme of partitioning data, multiple independent runs, statistical significance testing, confidence interval estimation, and comprehensive ablation analysis. The experimental results demonstrated that the complete framework achieved an accuracy of 92.0% with a Word Error Rate (WER) of 0.07, outperforming the baseline and intermediate voting strategies. Statistical analysis confirmed that the observed improvements were significant, while the ablation study verified the complementary contributions of MGPT segmentation, skeleton-based representation, WPG-guided structured decoding, and Gaussian-weighted temporal decision fusion. In addition, cross-signer evaluation indicated good generalization capability despite variations in signing styles.

Besides the current study, which is limited to the ISL-CSLRT dataset and does not involve real-time deployment or gloss-level recognition, the results give evidence that explicit temporal segmentation together with structured decoding is a way to significantly improve CSLR with weak supervision. Next research will be more on CSLR benchmark datasets, adoption of transformer-based temporal modeling and multi-modal rich representations, study of semi-supervised learning strategies for low-resource sign languages, and development of efficient real-time CSLR systems for assistive communication purposes.

Conflicts of Interest

The authors declare no competing interests.

Funding Statement

Nil

Acknowledgments

Paresh Chauhan conceptualized the study, designed the proposed flow, and Dineshkumar Vaghela supervised the overall research process. Mahesh Goyani contributed to the implementation of the proposed flow and validate the results. Udesang Jaliya briefly did the literature survey. All authors discussed the results, reviewed, and approved the final version of the manuscript.

References

[1] Necati Cihan Camgöz et al., "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 10023-10033, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Oscar Koller, "Quantitative Survey of the State of the Art in Sign Language Recognition," *arXiv preprint*, pp. 1-40, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Hao Zhou et al., "Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13009-13016, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yuecong Min et al., "Visual Alignment Constraint for Continuous Sign Language Recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11522-11531, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Dongxu Li et al., "Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, USA, pp. 1879-1889, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yao Du et al., "Full Transformer Network with Masking Future for Word-Level Sign Language Recognition," *Neurocomputing*, vol. 500, pp. 115-123, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Runpeng Cui, Hu Liu, and Changshui Zhang, "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880-1891, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera, "Sign Language Recognition: A Deep Survey," *Expert Systems with Applications*, vol. 164, pp. 1-89, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Songyao Jiang et al., "Skeleton Aware Multi-modal Sign Language Recognition," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, pp. 328-338, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ashish Vaswani et al., "Attention is all you Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 1-11, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Junfu Pu, Wengang Zhou, and Houqiang Li, "Iterative Alignment Network for Continuous Sign Language Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 4160-4169, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Hannah Bull et al., "Aligning Subtitles in Sign Language Videos," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11532-11541, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sijie Yang, Yuanjun Xiong, and Dahua Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 7444-7452, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] Zhe Cao et al., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Fan Zhang et al., "MediaPipe Hands: On-device Real-time Hand Tracking," *arXiv preprint*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Pan Xie et al., "Multi-Scale Local-Temporal Similarity Fusion for Continuous Sign Language Recognition," *Pattern Recognition*, vol. 136, pp. 1-33, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Borui Miao et al., "DC-BVM: Dual-Channel Information Fusion Network based on Voting Mechanism," *Biomedical Signal Processing and Control*, vol. 94, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Hezhen Hu et al., "SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11067-11076, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Lu Meng, and Ronghui Li, "An Attention-Enhanced Multi-Scale and Dual Sign Language Recognition Network Based on a Graph Convolution Network," *Sensors*, vol. 21, no. 4, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Weichao Zhao et al., "Self-Supervised Representation Learning With Spatial-Temporal Consistency for Sign Language Recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 4188-4201, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] R. Elakkiya, and B. Natarajan, "ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition," *Mendeley Data V1*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] M. Madhiarasan, and Partha Pratim Roy, "A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets," *arXiv preprint*, pp. 1-30, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mohammed Waleed Kadous, "Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language," *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pp. 1-10, 1996. [Google Scholar]

[24] C. Vogler, and D. Metaxas, "Parallel Hidden Markov Models for American Sign Language Recognition," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, Greece, pp. 116-122, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[25] Helen Cooper, Brian Holt, and Richard Bowden, *Sign Language Recognition*, Visual Analysis of Humans, pp. 539-562, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[26] Sarah Alyami, and Hamzah Luqma, "Swin-MSTP: Swin Transformer with Multi-Scale Temporal Perception for Continuous Sign Language Recognition," *Neurocomputing*, vol. 617, 2025. [CrossRef] [Google Scholar] [Publisher Link]