Optimizing sequential models through temporal landmark selection and normalization for sign language recognition

Sergio Esteban-Romero, Iván Martín-Fernández, Cristina Luna-Jiménez, Manuel Gil-Martín, Fernando Fernández-Martínez, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Esteban-Romero, Sergio, Iván Martín-Fernández, Cristina Luna-Jiménez, Manuel Gil-Martín, Fernando Fernández-Martínez, and Elisabeth André. 2026. "Optimizing sequential models through temporal landmark selection and normalization for sign language recognition." In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026), March 5-8, 2026, Marbella, Spain, volume 3*, edited by Ana Paula Rocha, Mattias Wahde, and H. Jaap van den Herik, 2499–2506. Setúbal: SciTePress. https://doi.org/10.5220/0014281100004052.

Optimizing Sequential Models through Temporal Landmark Selection and Normalization for Sign Language Recognition

Sergio Esteban-Romero[1][a], Iván Martín-Fernández[1][b] Cristina Luna-Jiménez[2][c], Manuel Gil-Martín[1][d], Fernando Fernández-Martínez[1][e] and Elisabeth André[2][f]

[1]*Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain*

[2]*Human-Centered Artificial Intelligence, University of Augsburg, Augsburg, Bayern, Germany*

Keywords: Isolated Sign Language Recognition, Body Structure Landmarks, Machine Learning, Human-Computer Interaction, Accessibility.

Abstract: This paper presents a comprehensive study of lightweight and efficient models for isolated sign language recognition that achieve consistent performance in two benchmark datasets, AVASAG100 and WLASL100. Our pipeline leverages ViTPose to extract keypoints from every video frame, which are then processed by Transformer-based architectures. We investigate the impact of keypoint normalization and centering across frames, as well as the role of a learnable weighted pooling mechanism compared to uniform averaging of encoder representations. Our findings show that normalization, particularly centering, is crucial for stable and effective learning. Furthermore, while learnable weighted pooling enhances performance in LSTM-based models, it provides limited benefit for Transformers, indicating that Transformer representations are inherently more robust. Moreover, an ablation study related to a simple yet effective frame selection strategy based on mean hand-confidence thresholds led to improved efficiency efficiency. On WLASL100, the proposed compact Transformer with only 2.2M parameters achieves strong results, 65.99% M-F1, 66.24% W-F1, and 67.83% Accuracy, demonstrating the effectiveness of our approach in balancing efficiency and performance.

1 INTRODUCTION

Recent advances in Human-Computer Interaction have significantly improved access to technology and communication opportunities for deaf individuals (Kuhn et al., 2025). From a socio-anthropological approach, the deaf community is recognized as a minority group with their own social and cultural characteristics (Rodríguez-Correa et al., 2023). Furthermore, in some scenarios such as healthcare, the presence of an interpreter, often a family member, is frequently required for patients to effectively communicate their needs and experiences. The situation becomes even more complex in mental health settings, where it is estimated that 1 in every 8 people in the world live with a mental disorder. For deaf individuals, seeking therapy may be particularly challenging, as they often feel uncomfortable sharing personal thoughts and emotions through an interpreter. These barriers highlight the critical role that accessible, direct communication technologies can play in supporting the mental health and well-being of the deaf community.

In this context, advances in Sign Language processing offer promising solutions. This field can be categorized into two primary paradigms. The first is Isolated Sign Language Recognition (ISLR), which focuses on identifying individual signs, or glosses, performed by a signer in an image or video segment. This approach treats each sign as a discrete unit, making it suitable for applications such as vocabulary learning, gesture-based interfaces, and sign lexicon building (Sarhan and Frintrop, 2023). The second paradigm is Continuous Sign Language Translation (CSLT), which aims to directly translate a continuous sequence of signed gestures into natural language

[a] https://orcid.org/0009-0008-6336-7877
[b] https://orcid.org/0009-0004-2769-9752
[c] https://orcid.org/0000-0001-5369-856X
[d] https://orcid.org/0000-0002-4285-6224
[e] https://orcid.org/0000-0003-3877-0089
[f] https://orcid.org/0000-0002-2367-162X

Esteban-Romero, S., Martín-Fernández, I., Luna-Jiménez, C., Gil-Martín, M., Fernández-Martínez, F. and André, E.
Optimizing Sequential Models through Temporal Landmark Selection and Normalization for Sign Language Recognition.
DOI: 10.5220/0014281100004052

In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026) - Volume 3*, pages 2499-2506
ISBN: 978-989-758-796-2; ISSN: 2184-433X

sentences from a video of the signer. Unlike ISLR, CSLT must handle temporal dependencies, coarticulation effects, and contextual information across multiple signs, making it a more complex task. However, overcoming these challenges enables direct applications in real-time communication and accessibility tools (Núñez-Marcos et al., 2023), improving the quality of interaction for deaf individuals.

While advancements in both ISLR and CSLT rely heavily on precise pose estimation, most current landmark-based ISLR models use all the landmarks detected by keypoint extractors. This practice may force the network to learn unnecessary or spurious dependencies, many of which are not actually relevant to the intended recognition task (Pu et al., 2025).

This work focuses specifically on ISLR and presents a comprehensive evaluation of lightweight, efficient models capable of delivering consistent performance across two distinct benchmark datasets. The main objectives of this study are the following:

1. Architecture comparison: To compare Long Short-Term Memory (LSTM) networks and lightweight Transformer encoder–decoder architectures, matched by parameter count, to identify which provides superior efficiency and accuracy for ISLR when processing body-landmark inputs.
2. Feature-group combination: To explore and validate strategies for combining available keypoints, determining whether this feature-selection procedure can further improve overall performance.

By addressing these objectives, this study aims to improve classification performance while minimizing both model size and inference latency, critical factors for seamless integration in real-time CSLT systems.

The remainder of this work is structured as follows, Section 2 explores the current state-of-the-art solutions for ISLR focused on landmark input based models. Section 3 describes the datasets used in this study. Section 4 explains the preprocessing that have been tested and the model architecture used. Section 5 reports the results obtained with a related discussion. Section 6 highlights the relevant insights gained throughout the experimentation process.

2 RELATED WORK

Sign Language recognition problems are typically approached using either RGB-based or pose-based methods, depending on whether the model processes raw images (RGB) or extracted skeleton data. This study specifically focuses on the latter, leveraging pose-based techniques. To generate a usable representation of a person's body configuration from RGB images or video, a landmark extractor model is used to locate specific keypoints. In ISLR, most models incorporate an encoder to produce a compressed representation of the input, followed by a classification head that identifies the gloss being represented.

State-of-the-art pose estimation includes CNN-based architectures like MediaPipe, and OpenPose (Cao et al., 2019), which utilizes part affinity fields to detect 135 keypoints. In contrast, ViTPose (Xu et al., 2022) employs a Vision Transformer backbone to generate 133 keypoints following the COCO-WholeBody format (Jin et al., 2020). Recent years have seen a surge of transformer-based models for ISLR covering both video-based models and body keypoint modalities. SPOTER (Boháček and Hrúz, 2022) applies a Transformer encoder-decoder to flattened landmark sequences (54 joints $\times$ 2). The method utilizes extensive spatial augmentation and a specific class-query decoding mechanism, resulting in a reported 63.18% accuracy on the WLASL100 benchmark. Adopting a self-supervised approach, SignBERT (Hu et al., 2021) treats hand poses as visual tokens and employs BERT-style masked reconstruction. By integrating spectral graph convolutional networks (GCNs) with positional embeddings, the framework achieved 76.36% accuracy, later improved to 79.84% in its enhanced version. Recent skeleton-based approaches have shown significant improvements on the WLASL100 dataset. Sign2Pose (Eunice et al., 2023) achieved 80.9% accuracy by combining Vision API landmarks with refined augmentation techniques. Improving on realistic representation, Pu et al. (Pu et al., 2025) introduced kinematic hand pose rectification and feature isolation to decouple body dependencies, reaching 86.50%. Currently, UniSign (Li et al., 2025) holds the state-of-the-art (92.24%) by independently encoding and concatenating features from the hands, face, and upper body, with optional RGB integration.

3 DATASETS

This section describes the datasets used, detailing their characteristics and relevance to the ISLR tasks.

3.1 AVASAG100

The AVASAG dataset (Bernhard et al., 2022; Nunnari et al., 2021) was created to encompass everyday travel-related scenarios. It contains German text and gloss annotations to support research in both ISLR and CSLT. In total, it comprises 312 sentence videos

recorded at a resolution of 1920×1080 pixels and 60 fps, with an overall duration of 96.05 minutes. Glosses were extracted from sentence-level annotations by defining the start time of each gloss as the end time of the preceding one, ensuring that transitional information is preserved (Konrad et al., 2022). This segmentation approach used the NOVA platform[1].

This study focuses exclusively on the ISLR material, utilizing a subset where the 100 most frequent glosses are selected, following a procedure analogous to the creation of the Word-Level American Sign Language (WLASL) (Li et al., 2020) dataset. The resulting dataset contains 4,336 gloss videos in total, partitioned into 2,596 for training, 849 for validation, and 890 for testing, keeping the class distribution consistent across all splits. Gloss repetition frequencies range from 470 for the most frequent to 14 for the least frequent ($\mu = 43.35$; $\sigma = 54.19$). All AVASAG100 videos were recorded with a single signer, positioned consistently in the frame and filmed in uniform background and lighting conditions.

This dataset was selected because of its clean, high-quality videos, which were recorded under expert supervision. This level of control contrasts with other large-scale public datasets, where data collection relies on metadata and lacks expert review.

3.2 WLASL100

The WLASL dataset (Li et al., 2020) is the largest-public American Sign Language dataset supporting word-level sign recognition research. It contains videos depicting more than 2,000 distinct words performed by over 100 signers. For the present study, we focus on the WLASL100 subset, which includes the 100 most frequent glosses, resulting in 2,038 videos signed by 97 different individuals. These videos were sourced from multiple educational sign language websites, ensuring their accuracy, and also from YouTube, where video titles clearly correspond to the depicted gloss. This set contains a wide variety of signs, such as apple, forget, medicine, or mother among others. The dataset is partitioned into 1,442 training examples, 338 for validation, and 258 for test. Gloss frequency ranges from 40 for the most common to 18 for the least frequent ($\mu = 20.4$; $\sigma = 3.28$).

This dataset was selected for its substantial signer diversity and its wide variety of backgrounds, recording setups, and conditions. Evaluating on WLASL100 shows how well the proposed solutions generalize to varied real-world conditions.

[1]NOVA is an open-source annotation tool for multimodal data, commonly used in sign language corpus projects. See: https://github.com/hcmlab/nova

4 METHODOLOGY

In this section, we discuss the preprocessing techniques and algorithms applied to our input data with a description of the whole proposed architecture, which has been illustrated in Figure 2.

4.1 Landmark Feature Extractor

To address a key limitation identified in a previous work (Esteban-Romero et al., 2025), where MediaPipe (Zhang et al., 2020) struggled to accurately capture hand landmarks in the AVASAG100 dataset, we adopted ViTPose (Xu et al., 2022). The difficulty with MediaPipe was primarily attributed to the presence of gloves worn by the signer, which featured different colors when used in motion capture and hindered effective hand framing in the images. Specifically, we utilized the ViTPose-H model[2] with a resolution of 256x192, which comprises 637M parameters. This model outputs 133 landmarks covering the body, hands, pose, and feet, formatted according to the COCO-WholeBody standard (Jin et al., 2020). Each landmark is described by its x and y coordinates, accompanied by an acquisition confidence value. The ViTPose-H model was selected based on its demonstrated capabilities within the HaMeR framework (Pavlakos et al., 2024), where it successfully detected hands on a vast number of images of different nature for subsequent 3D hand mesh reconstruction. To ensure spatial invariance across varying video resolutions, all landmark coordinates are linearly normalized to the range $[-1, 1]$ and horizontally centered relative to the nose keypoint.

4.2 Frame Selection by Hand-Confidence Filtering

In our analysis, we utilize the landmarks confidence scores output by the ViTPose model to temporally localize the target gloss sign within each video. We observed that in most sign language videos, the initial and final frames often lack of movement, and notably, the signer's hands, which are unequivocally the most informative body group, may not be visible in the image. To address this issue, we propose a straightforward preprocessing algorithm that evaluates the mean confidence of ViTPose hand keypoints on a frame-by-frame basis, selecting the time intervals most likely to contain the target action for retention and subsequent processing. The algorithm steps would be as follows:

[2]https://github.com/ViTAE-Transformer/ViTPose

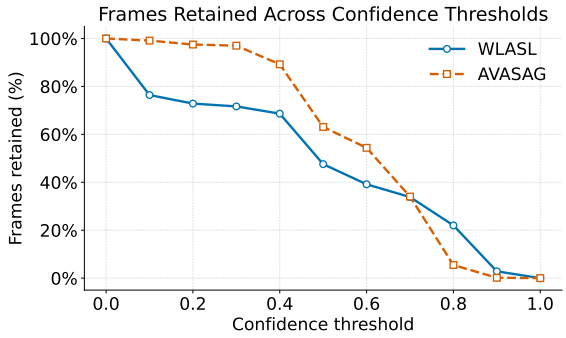

Figure 1: Effect of confidence threshold filtering on frame retention for the AVASAG100 andWLASL100 datasets.

1. Compute the *mean hand confidence* for every video frame, defined as the average ViTPose confidence score of the detected hand landmarks within that individual frame.
2. Select frames whose mean hand confidence exceeds a fixed threshold (default set to 0.0), ensuring that only frames with reliable hand detections are kept.
3. If no frames satisfy this criterion, such as in low-quality or partially occluded sequences, the method switches to a dynamic fallback, using the 75th percentile of the confidence distribution as a relaxed threshold.
4. To guarantee sufficient temporal coverage, it enforces a minimum number of frames (default 15). If the selection is still too small, the algorithm progressively includes frames with the highest confidence values, lowering the fallback percentile to 50% if necessary.
5. Finally, the chosen frame indices are sorted to reconstruct the processed video segment.

This adaptive procedure yields a robust set of frames that effectively captures the time intervals during which hand-driven actions are detected with the highest confidence, even under challenging imaging conditions. Figure 1 illustrates the relationship between the confidence threshold and the proportion of frames retained across the WLASL100 and AVASAG100 datasets. As the confidence threshold increases, the proportion of retained frames decreases consistently for both datasets, reflecting the stricter filtering of frames with low-confidence hand landmark detections. We hypothesize that employing threshold values above 0.5 for both datasets might lead to suboptimal performances, since the fallback mechanisms are activated for a considerable number of videos, potentially leading to a measurable loss in terms of temporal resolution and coherence.

4.3 Model Architecture

In this study, we work with landmarks represented by their spatial location (i.e. cartesian x- and y-coordinates) and their temporal dynamics across video frames. To establish a strong model baseline, we employ LSTM networks, which have been proven effective in capturing temporal dependencies and have seen extensive application in sign language recognition (Mittal et al., 2019). In addition, Transformer-based models have also demonstrated remarkable effectiveness in sequence modeling across diverse domains, owing to their attention mechanisms, including recent advances in sign language recognition (Boháček and Hrúz, 2022).

By incorporating both LSTM and transformer encoders, we aim to comprehensively assess temporal modeling approaches for landmark-based ISLR. While recognizing the architectural differences between LSTM and transformer-based systems, we established comparable parameter counts for both encoder types to ensure a fair comparison. Although it could be argued that this may lead to suboptimal architectures, in this work, we have prioritized a higher-level analysis that is also more comprehensive and offer greater generalizability. By maintaining similar model complexity, the insights gained are more likely to remain relevant and informative in further research.

Figure 2 illustrates the overall architecture adopted in this study. The pipeline begins by processing each video with a landmark extraction model that generates keypoint coordinates for distinct body regions, including body joints, hands, and facial contours. In our case, we utilized ViTPose, which returns the x and y cartesian coordinates for each landmark in every frame, along with a confidence value measuring the certainty of the prediction. Then, the described trajectories throughout the video are encoded using either an LSTM or a Transformer Encoder, resulting in a sequence of representations of equal length as the input. To aggregate temporal information and compress the encoder output, a pooling operation is applied to generate a final representation for each video. This representation is subsequently passed through a linear layer, which produces the logits and maps them to one of the gloss classes in the dataset dictionary.

In addition, this work examines the impact of different pooling strategies applied to the encoder output, specifically comparing average pooling to a learnable weighted pooling mechanism implemented via a linear layer. This simple mechanism minimally increases the model's parameter count but can enrich the final embeddings for classification.

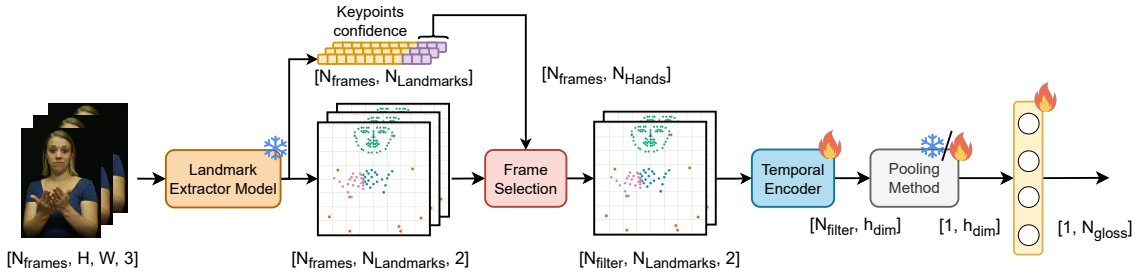

Figure 2: The proposed architecture to perform ISLR. $N_{Landmarks}$ refers to the amount of landmarks generated by the landmark extractor model employed. The Frame Selection model uses confidence of hand-keypoint recognition to filter frames. The Temporal Encoder can be either a LSTM or a Transformer encoder architecture. Pooling method can be done either as the average of the output representation or through a learnable weight vector. The final stage performs classification using a linear layer mapping to the predicted gloss. The flame indicates elements finetuned and the snowflake indicates elements that remain frozen throughout the training process.

4.3.1 LSTM Encoder

The LSTM model aims to capture mainly temporal dependencies within the landmark sequences. In particular, we implemented a Bidirectional LSTM (Bi-LSTM) architecture, that processes the sequence in both forward and backward directions to better capture contextual information. We used a compact configuration of Bi-LSTM to ensure model capacity while maintaining computational efficiency, and without extensive hyperparameter tuning. The encoder is a Bi-LSTM model (2.62M parameters) configured with 2 layers and a hidden dimension of $h_{lstm} = 256$.

Note that Bi-LSTMs produce separate representations for the forward and backward temporal dependencies. These two representations are concatenated (stacked), effectively doubling the output dimensionality (i.e. $h_{dim} = 2 * h_{lstm}$) as illustrated in Figure 2.

4.4 Transformer Encoder

Following the methodology of SPOTER (Boháček and Hrúz, 2022; Luna-Jiménez et al., 2023), we employ a transformer encoder architecture to obtain intermediate representations that capture the temporal dynamics of the extracted landmarks. To ensure consistency across varying landmark configurations, we incorporated a linear projection layer that maps the input landmarks into a fixed embedding dimensionality. Moreover, by standardizing the input representation, the architecture avoids potential constraints on the multi-head attention mechanism, as the number of attention heads must divide the embedding dimensionality evenly. Beyond dimensional compatibility, this linear layer enhances flexibility, promotes architectural stability, and supports scalability by decoupling the variability of input landmark sets from the requirements of the downstream transformer encoder.

We also include positional encoding, introducing deterministic sinusoidal patterns to represent sequence order in transformer architectures, enabling the model to incorporate information about the relative and absolute positions of tokens without relying on recurrence. The encoding is precomputed up to a maximum sequence length and added element-wise to the input embeddings before the subsequent transformer layers, with dropout to mitigate overfitting.

The encoder is a Transformer model (2.2M parameters) configured with 4 layers, 8 heads, $d_{model} = 256$, $dim_{fw} = 512$, and a dropout rate of 0.1. Note that in Figure 2, $h_{dim} = d_{model}$.

5 RESULTS

This section provides an overview of the experimental conditions, including discussions and insights.

5.1 Experimental Conditions

Two different training scenarios were considered, each tailored to one of the datasets under analysis based on observed convergence patterns within their respective validation sets. The models were trained for 50 and 100 epochs in AVASAG100 and WLASL100, respectively. This difference is justified by the greater similarity and uniformity in AVASAG100 videos, where the signer consistently occupies the same position, in contrast to the higher variability of signers in the WLASL100 dataset. The remainder of the hyperparameters are shared in both scenarios and no exhaustive hyperparameter tuning is performed. The learning rate is set to 10^{-4} with Adam optimizer and the batch size is set to 16. The experiments were carried out on a NVIDIA GeForce RTX3090 24GB and a Tesla V100 32GB.

5.2 AVASAG100

In this section, we report the results obtained for the models on the AVASAG100 test partition trained with training and validation splits. The validation split was used to determine a reasonable number of epochs and a learning rate to ensure the training convergence.

Table 1 presents a comparative analysis between LSTM and Transformer-based architectures under different pooling strategies and normalization settings, using all available landmarks. Both models achieve their highest performance when weighted pooling is combined with normalization: LSTM reaches M-F1=84.64% W-F1=88.77% similar to the Transformer with M-F1=84.87%, W-F1=88.01%. In contrast, when average pooling is applied, LSTM performance drops significantly, whereas the Transformer maintains performance comparable to the weighted pooling scenario. This indicates that Transformer-generated representations are robust and informative regardless of the pooling method, in contrast to LSTMs, which benefit substantially from learnable weighting pooling. Furthermore, omitting normalization leads to a pronounced performance deterioration across both architectures, underscoring the critical role of normalization with centering for achieving stable and effective learning in this scenario. Taken together, the results show that the weighted pooling with normalization provides the most effective configuration for both architectures.

Table 1: Comparison of LSTM vs. Transformer with a similar parameter count in the AVASAG100 test set.

Model	Pooling	Norm.	M-F1	W-F1	Acc
LSTM	Weight.	✓	84.64	88.77	88.87
		✗	54.29	61.25	63.48
	Avg.	✓	73.69	80.11	80.79
		✗	29.37	40.86	46.07
Transf.	Weight.	✓	84.87	88.01	88.43
		✗	56.92	65.89	66.97
	Avg.	✓	83.23	87.38	87.75
		✗	59.20	66.12	67.86

5.2.1 Hand-Confidence Filtering Analysis

In this section, we evaluate the influence of selecting frames where the signer's hands are more clearly visible. Table 2 shows that the LSTM model achieves its best performance at an average hand-confidence threshold of 0.1, reaching M-F1=86.93% and W-F1=89.62%. However, increasing the threshold to 0.6 progressively decreases the performance, dropping to M-F1=74.13% and W-F1=79.12%. The Transformer model exhibits a similar trend, achieving its best performance at the 0.1 threshold, with a progressive decline that reaches its minimum at 0.6. These results underscore that filtering out frames with low hand-confidence scores can improve model performance, although the gains are not statistically significant.

Table 2: Transformer and LSTM results depending on the threshold for the mean confidence of the hand in AVASAG100.

Model	Avg. Hands Confidence	M-F1	W-F1	Acc
LSTM	0.0	84.64	88.77	88.87
	0.1	86.78	89.16	89.21
	0.2	83.67	87.70	88.20
	0.3	85.94	88.99	89.21
	0.4	80.56	84.71	85.17
	0.5	76.79	80.57	81.12
	0.6	74.13	79.12	79.77
Transf.	0.0	84.87	88.01	88.43
	0.1	84.99	88.77	89.10
	0.2	83.20	87.09	87.53
	0.3	83.92	87.84	88.31
	0.4	79.73	84.09	84.61
	0.5	75.69	80.97	81.46
	0.6	74.42	79.56	80.34

5.3 WLASL100

In this section, we discuss the results obtained for the WLASL100 dataset, which is characterized by significant variability across signers and recording conditions. Table 3 offers a comparison between the architectures, pooling strategies, and normalization methods evaluated in this study. For LSTMs, the best results are achieved using learnable weighted pooling combined with normalization, whereas the compact Transformer encoder performs best with average pooling and normalization. Overall, learnable weighted pooling proves effective only for the LSTM-generated representations, while its impact is limited when applied to Transformer embeddings. This highlights the Transformer's ability to generate robust and equally informative embeddings regardless of the pooling method. Regarding normalization, it consistently improves performance across all scenarios, although Transformers demonstrate higher robustness to raw input data, likely reflecting their capacity to better handle variability.

5.3.1 Hand-Confidence Filtering Analysis

In this scenario, we use all available landmarks and apply the proposed frame selection algorithm.

Table 3: Comparison of LSTM vs. Transformer with similar parameter count in WLASL100 test set.

Model	Pooling	Norm.	M-F1	W-F1	Acc
LSTM	Weight.	✓	57.10	57.19	58.14
		✗	31.41	30.79	34.49
	Avg.	✓	47.47	47.42	49.61
		✗	11.11	10.82	13.96
Trans.	Weight.	✓	54.19	54.18	56.59
		✗	41.36	41.11	42.63
	Avg.	✓	57.92	57.53	59.69
		✗	38.92	37.91	41.08

As shown in Table 4, varying the mean hand confidence threshold substantially impacts model performance. For LSTMs, with no filtering (threshold = 0.0), the system achieves M-F1=57.10%, W-F1=57.19%. Increasing the threshold to 0.2 yields peak performance at M-F1=60.62%, W-F1=60.22%, but further increases result in diminishing returns, suggesting that relevant information is also being removed. Transformers exhibit a similar but even more pronounced trend, starting at M-F1=54.19% and W-F1=54.18% for a threshold of 0.0, they reach their best performance at a threshold of 0.4 with M-F1=65.99% and W-F1=66.24%, and a top-1 accuracy of 67.83%. These results underscore that filtering out low-confidence frames improves performance, validating the effectiveness of the confidence-based selection mechanism, particularly for low quality videos that are more prone to temporal or spatial resolution issues, thereby obtaining worse keypoint predictions.

Table 4: Transformer and LSTM results depending on the threshold for the mean confidence of the hand in WLASL100.

Model	Avg. Hands Confidence	M-F1	W-F1	Acc
LSTM	0.0	57.10	57.19	58.14
	0.1	55.24	54.72	56.98
	0.2	60.62	60.22	61.63
	0.3	59.48	59.13	60.47
	0.4	57.42	57.69	59.69
	0.5	50.81	50.86	53.49
Trans.	0.0	54.19	54.18	56.59
	0.1	60.99	60.51	62.40
	0.2	63.18	62.70	65.11
	0.3	63.61	63.21	65.12
	0.4	65.99	66.24	67.83
	0.5	57.32	57.00	58.91

Compared to the results obtained for AVASAG100, where barely no gain is obtained, our findings indicate that the frame selection mechanism is particularly effective when applied to noisier data. This approach allows a simple yet efficient means of ensuring that only the highest quality frames are used to model the spatio-temporal dynamics of the gloss under analysis. Notably, the best result reported for the Transformer model in Table 4 is encouraging: even in the absence of exhaustive hyperparameter tuning, performance approximates leading state-of-the-art solutions like SignBert (Acc=76.36%) (Hu et al., 2021). This idea suggests that further hyperparameter optimization might lead to even better results, especially considering that we only used 2.2M parameters, significantly fewer than most recent state-of-the-art models.

6 CONCLUSIONS AND FUTURE WORK

This study addressed ISLR using two different datasets: AVASAG100, containing high-quality, professionally recorded videos of a single German Sign Language signer, and WLASL100, which comprises internet and educational videos in American Sign Language without expert curation. Our proposed pipeline employs ViTPose to extract body landmarks frame-by-frame, followed by a Transformer encoder to model temporal relationships and generate representations. These are aggregated through either average or learnable weighted pooling before a linear classification layer predicts the gloss label.

Results show that normalization and centering are crucial for stable and effective learning, while the learnable weighted pooling benefits LSTM models more than Transformer encoders, suggesting that latter representations are more robust. Additionally, we introduced a simple yet efficient frame selection mechanism based on mean hand confidence thresholds. This approach effectively filters out low-confidence frames, significantly improving performance on the noisier WLASL100 dataset (achieving M-F1=65.99%, W-F1=66.24%, and Acc=67.83%), while yielding little gain on the cleaner AVASAG100 dataset. This demonstrates the utility of confidence-based frame selection in handling low-quality videos.

For future work, exploring strategies to process and integrate the most informative body groups could enhance recognition performance. Moreover, training was conducted with a standardized number of epochs and hyperparameter settings chosen for practical convergence, rather than optimal configurations per model or scenario. This leaves potential for further performance gains through more exhaustive tuning and architecture adaptation.

ACKNOWLEDGEMENTS

S.E.-R.'s research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). The research of I.M.-F. was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). This work was funded by the Spanish Ministry of Science and Innovation through the projects TRUSTBOOST (PID2023-150584OB-C21 and PID2023-150584OB-C22), and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

REFERENCES

Bernhard, L., Nunnari, F., Unger, A., Bauerdiek, J., Dold, C., Hauck, M., Stricker, A., Baur, T., Heimerl, A., André, E., et al. (2022). Towards automated sign language production: A pipeline for creating inclusive virtual humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 260–268.

Boháček, M. and Hrúz, M. (2022). Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 182–191.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

Esteban-Romero, S., Luna-Jiménez, C., Gil-Martín, M., Fernández-Martínez, F., and Andre, E. (2025). Llm-driven multimodal video-text fusion for isolated sign language recognition. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, IVA Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Eunice, J., J, A., Sei, Y., and Hemanth, D. J. (2023). Sign2pose: A pose-based approach for gloss prediction using a transformer model. *Sensors*, 23(5).

Hu, H., Zhao, W., Zhou, W., Wang, Y., and Li, H. (2021). Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096.

Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., and Luo, P. (2020). Whole-body human pose estimation in the wild. *CoRR*, abs/2007.11858.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2022). Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen.

Kuhn, K., Kersken, V., and Zimmermann, G. (2025). Communication access real-time translation through collaborative correction of automatic speech recognition. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Li, D., Rodriguez, C., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., and Li, H. (2025). Uni-sign: Toward unified sign language understanding at scale.

Luna-Jiménez, C., Gil-Martín, M., Kleinlein, R., San-Segundo, R., and Fernández-Martínez, F. (2023). Interpreting sign language recognition using transformers and mediapipe landmarks. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 373–377, New York, NY, USA. Association for Computing Machinery.

Mittal, A., Kumar, P., Roy, P. P., Balasubramanian, R., and Chaudhuri, B. B. (2019). A modified lstm model for continuous sign language recognition using leap motion. *IEEE Sensors Journal*, 19(16):7056–7063.

Nunnari, F., Bauerdiek, J., Bernhard, L., Espana-Bonet, C., Jäger, C., Unger, A., Waldow, K., Wecker, S., André, E., Busemann, S., Dold, C., Fuhrmann, A., Gebhard, P., Hamidullah, Y., Hauck, M., Kossel, Y., Misiak, M., Wallach, D., and Stricker, A. (2021). Avasag: A german sign language translation system for public services. In *Proceedings of Machine Translation Summit XVIII*, pages 43–48.

Núñez-Marcos, A., de Viñaspre, O. P., and Labaka, G. (2023). A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., and Malik, J. (2024). Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836.

Pu, M., Lim, M. K., and Chong, C. Y. (2025). Siformer: Feature-isolated transformer for efficient skeleton-based sign language recognition.

Rodríguez-Correa, P. A., Valencia-Arias, A., Patiño-Toro, O. N., Oblitas Díaz, Y., and Teodori De la Puente, R. (2023). Benefits and development of assistive technologies for deaf people's communication: A systematic review. *Frontiers in Education*, Volume 8 - 2023.

Sarhan, N. and Frintrop, S. (2023). Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3210–3219.

Xu, Y., Zhang, J., Zhang, Q., and Tao, D. (2022). Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.