A Lightweight Solution for Pose-Based Recognition for Isolated Spanish Sign Language Using Recurrent Models

Gerardo León-Quintana, José Salas-Cáceres and Javier Lorenzo-Navarro
Universidad de Las Palmas de Gran Canaria, Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería (SIANI), Las Palmas de Gran Canaria, Spain
gerardo.leon101@alu.ulpgc.es,{jose.salas, javier.lorenzo}@ulpgc.es

Keywords: Human-Machine Interaction, Biometry, Machine Learning, Sign Language Recognition.

Abstract: This work addresses the problem of Isolated Sign Language Recognition (ISLR) in Spanish Sign Language (LSE) from a pose-based perspective. The proposed approach relies on 3D landmark extraction using Google's MediaPipe framework to obtain face, hand, and upper-body keypoints, which are then normalized and transformed into spatial–temporal feature sequences. Two temporal alignment strategies, *average sampling* and *max-length padding*, were implemented to ensure uniform input dimensions across samples. Bidirectional recurrent neural networks (Bi-LSTM and Bi-GRU) were evaluated to capture the temporal dependencies inherent to signing. Experimental results on the LSE-Health-UVigo dataset show that the Bi-LSTM architecture combined with the Focal Loss function ($\gamma = 3$) achieved the highest performance, reaching 79.8% unweighted accuracy. The proposed model has an average response time of approximately 1 ms, making it suitable for deployment in real-time scenarios. These results highlight the effectiveness of pose-based recurrent architectures for ISLR and demonstrate the potential of lightweight models for robust sign language understanding.

1 INTRODUCTION

Human–machine interaction (HMI) has undergone significant transformations in recent years, driven by advances in technological fields such as machine learning, natural language processing (NLP), and computer vision (Lv et al., 2022). These developments have enabled systems capable of performing complex tasks, including speech recognition, biometric identification, and object detection.

Despite these achievements, a significant gap remains in accommodating users who communicate primarily through sign language. Most current systems depend on spoken or written language as the principal medium of interaction, providing limited or no support for gestural communication.

This absence of automated sign language recognition (SLR) mechanisms not only restricts the accessibility of technological tools for the Deaf communities, but also underscores the need to develop technological solutions that can enhance communication, accessibility, and social integration for deaf and hard-of-hearing individuals.

Addressing this challenge requires a comprehensive understanding of sign languages, which are fully developed linguistic systems with their own grammar and syntax. Unlike spoken languages, they are visual–gestural and rely on manual movements, facial expressions, body posture, and spatial structure to convey meaning. These visual features are classified into: manual features and non-manual features (Sutton-Spence and Woll, 1999). The former relates to the hands movements and shapes while the latter refers to other body movements such as facial expressions, mouthing or the upper body posture. Traditionally the deep learning approaches employ only the manual features (Cooper et al., 2012), however, recent works (Saunders et al., 2021; Tarrés et al., 2023), have successfully employ both type to achieve good results.

Key linguistic units in sign languages include the gloss, which provides a written representation of a sign, usually expressed as a single word or short phrase capturing its conceptual meaning (Niu and Mak, 2020). Rather than serving as a direct translation into spoken language, glosses function as symbolic annotations that preserve the linguistic structure of signing. As such, they establish a link between the visual–gestural modality of sign languages and their textual representation, playing a fundamental role in the creation of annotated corpora and in the training of SLR systems.

León-Quintana, G., Salas-Cáceres, J. and Lorenzo-Navarro, J.
A Lightweight Solution for Pose-Based Recognition for Isolated Spanish Sign Language Using Recurrent Models.

In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026) - Volume 5*, pages 4356-4363
ISBN: 978-989-758-796-2; ISSN: 2184-433X

Sign languages have evolved naturally within Deaf communities, resulting in distinct national and regional variants with their own grammatical and semantic structures. This explains the absence of a universal sign language and underscores the linguistic independence of systems such as American Sign Language (ASL) and British Sign Language (BSL). Acknowledging this diversity is essential for developing of effective and linguistically grounded SLR technologies.

In addition to linguistic considerations, system latency is a critical factor in human–machine interaction (HMI). System response time (SRT) defines the delay between user input and system output, and values exceeding approximately 100 ms negatively impact perceived responsiveness and user satisfaction (Attig et al., 2017). Consequently, lightweight models with low inference latency are required to enable natural, fluid interaction and to support practical deployment in real-time HMI scenarios (Dautenhahn, 1995).

2 RELATED WORK

Visual sign language understanding typically involves three tasks: SLR, Sign Language Translation (SLT), and Sign Language Production (SLP) (Tarrés et al., 2023). SLR extract glosses corresponding to each sign performed by an individual performer (Coster et al., 2021) whereas SLT captures the overall meaning of a whole set of signs, often bypassing the gloss level and directly generating the corresponding sentence (Camgoz et al., 2021). Finally, SLP follows the reverse process, generating sign poses from either glosses or textual sentences (Saunders et al., 2021).

These tasks can be further categorized based on the temporal context of their features, leading to either static or dynamic approaches. The former focus on particular instants utilizing the information of a single pose, as seen in (Laines et al., 2023) where the authors predicted the gloss translation based on a single photograph of the pose. In contrast, dynamic solutions exploit temporal information leveraging methods such as Long Short-Term Memory (LSTM) layers (Fan et al., 2024; Wang et al., 2025) or transformers-based architectures (Saunders et al., 2021).

This static–dynamic distinction not only depends on the available resources but also reflects the intrinsic nature of the signs being analyzed. Static signs are characterized by fixed hand shapes and upper-body positions, typically used to represent letters, numbers, or other isolated symbols. In contrast, dynamic signs involve movement and encompass both manual motions and non-manual components such as facial expressions or changes in body posture. These temporal variations convey complete lexical meanings and constitute the majority of the vocabulary in sign languages. The coexistence of static and dynamic signs highlights the inherently multimodal and temporally structured nature of sign languages.

SLR, which is the task this work is focused on, is generally divided into two main approaches: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR) (Al Abdullah et al., 2024).

ISLR focuses on recognizing individual signs with clearly defined temporal boundaries. Each video corresponds to a single sign or gloss, which makes the task suitable for supervised learning. Most ISLR systems use convolutional, recurrent, or hybrid neural network architectures to learn spatial and temporal features from short video clips. For instance, Sarhan and Frintrop (Sarhan and Frintrop, 2023) present a comprehensive survey of isolated SLR methods, highlighting models such as 3D-CNNs and Bidirectional-LSTMs (Bi-LSTM) trained on short video segments to classify individual signs with high accuracy.

CSLR, in contrast, aims to recognize signs within continuous signing sequences, where gestures appear without explicit boundaries. This task is considerably more complex because the model must identify where one sign ends and another begins while maintaining linguistic coherence. As an example, (Alyami et al., 2024) review CSLR systems that combine convolutional and recurrent networks with alignment mechanisms like CTC to process long, unsegmented signing videos, addressing challenges such as coarticulation and movement epenthesis.

It is important to note that this distinction differs from the static-dynamic categorization, as a single sign can be composed of multiple moves, making a ISLR approach inherently dynamic.

As in many other machine learning applications, a major limitation in SLR is the limited availability of large, annotated, and publicly available datasets. Early datasets often contained a small number of isolated signs collected in controlled environments or using specialized sensors such as data gloves. In recent years, datasets like RWTH-PHOENIX-Weather 2014 (Koller et al., 2015), WLASL (Li et al., 2020) and BSL-1K (Albanie et al., 2021) have expanded the amount and diversity of available data. These include RGB video recordings, gloss annotations, and samples from multiple signers, enabling more realistic and generalizable recognition systems.

A particular issue in SLR is the gloss annotation of the different gestures, which is essential for training

translation models but is often absent in some public datasets such as the How2Sign dataset (Duarte et al., 2021) or the one introduced in (Yuan et al., 2019). This limitation arises from the high cost of providing the gloss translation of each sentence, which require very specialized personnel performing a very time-consuming activity.

Finally, SLR framework development is hindered by the diversity of sign languages which all are fully developed linguistic systems that rely on visual–gestural communication. All of this results in the absence of a "global" sign language, which not only hinders the possibility of conducting cross-dataset experiments but also underscores the need for technologies that facilitate communication between deaf communities across different cultures.

In this work, we present a lightweight Bi-LSTM model designed to perform dynamic ISLR for Spanish Sign Language (Lengua de Signos Española, LSE) using the LSE-Health-UVigo dataset (Alba-Castro et al., 2023), employing both manual and non-manual features.

3 DATASET

The LSE-Health-UVigo dataset, used in this study, contains approximately 11 hours of content and more than 270 videos of 11 Spanish Sign Language interpreters covering a variety of medical topics and explaining different health conditions. The recordings were captured under consistent lighting conditions and frame rates, ensuring uniform image quality across all samples. In total the database contains recordings of 105 different glosses. Figure 1 presents some representative frames that illustrate the overall visual quality of the dataset.

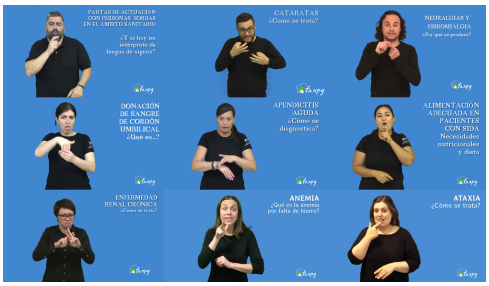

Figure 1: Representative frames from the LSE-Health-UVigo dataset.

However, many of these glosses exhibit very low usage frequencies or appear only in a limited number of videos performed by specific interpreters. This imbalance introduces considerable variability across classes and restricts the representativeness of several signs within the dataset. To mitigate this issue, a filtering criterion was established to retain only those gestures that met a minimum level of frequency and diversity across interpreters. Specifically, a gesture was required to be performed by at least three different interpreters, appear at least five times per interpreter, and accumulate a minimum of 60 total samples across all interpreters.

The filtering process was implemented using a Bag-of-Words (BoW) approach inspired by NLP methodologies. In this framework, each transcription was treated as a document, and a vocabulary was constructed from all annotated glosses, resulting in a 105-dimensional vector representation for each transcription. Each dimension corresponds to a unique gloss, and its value represents the number of occurrences of that gloss in the transcription. These vectors were subsequently aggregated by interpreter to identify which glosses were used by each signer throughout the entire set of videos. Applying the filtering to this aggregated representation lead to a final subset of 43 sign classes that fulfilled the established conditions.

Figure 2 shows the class distribution of the remaining signs. As observed, the resulting dataset is unbalanced, with certain classes appearing less than half as frequently as others.

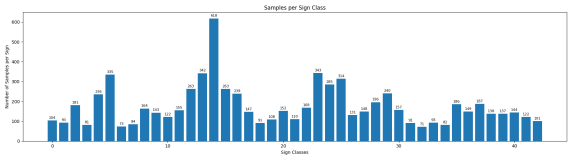

Figure 2: Histogram showing the distribution of signs remaining after data filtering.

This substantial reduction in the number of glosses was expected, given the medical nature of the dataset. Most videos focus on diseases affecting specific organs and are often interpreted by the same signer across multiple segments. Consequently, certain glosses became overrepresented by a single interpreter, while others lacked sufficient variability or frequency to be included in the final selection.

4 METHODOLOGY

This section outlines the data extraction pipeline and introduces the proposed dynamic architecture. It also presents the setup adopted during the experiments and the metrics used to evaluate the different results.

The proposed method is designed to recognize

isolated signs in LSE from RGB videos using a sequence of preprocessing, model training, and post-processing. The process begins with the extraction of visual features using pose estimation techniques, followed by the use of a normalization and temporal alignment. These features are then used to train recurrent neural network models that learn the temporal and spatial dependencies of each sign. And a final classification stage that assigns the sign to each video. The performance of the proposed method is evaluated using standard recognition metrics. The final classification stage assigns the sign to each video. The performance of the proposed method is evaluated using standard recognition metrics

The Figure 3 illustrates the complete processing pipeline. Starting from raw RGB video frames, MediaPipe extracts face, upper-body, and hand landmarks that describe the spatial configuration of each signer. These features are normalized to account for differences in body proportions and converted into structured vectors that preserve both spatial and temporal relationships. The resulting sequences are then processed by a dynamic model, which learns the temporal evolution of the gestures. Finally, the model generates a probability distribution over all sign classes, and the system assigns the label with the highest confidence score, completing the ISLR process.

4.1 Feature Extraction

For each video frame in the LSE-Health-UVigo dataset, 3D coordinates were extracted using Google's MediaPipe framework (Lugaresi et al., 2019). Two different models from this framework were employed: the Pose Landmark Detection model, which detects landmarks across the entire body, including the face; and the Hand Landmark Detection model, which captures the manual features of both hands.

Both of these models operate in 3D, meaning that each detected point includes not only its image-plane coordinates but also a depth value, allowing the estimation of spatial relationships, such as determining when the signer's hand is positioned in front of their face. From the resulting landmarks, a reduced subset of relevant keypoints was selected, focusing on the upper body as well as the most significant regions of the face and hands. An overview of the selected keypoints can be seen in Table 1.

To ensure consistency across samples, all coordinates were normalized by subtracting a central body reference point and scaling by the torso length. This normalization minimizes variations caused by differences in signers' body proportions. The resulting coordinates were used to compute inter-joint distances, capturing spatial relationships between hands, arms, and facial features.

Table 1: Summary of MediaPipe landmark models and their landmark examples. Fingers MCP refer to the metacarpophalangeal joints, commonly known as the knuckles.

Model	# Landmarks	# Used	Landmarks kept
Pose	33	3	Nose, shoulders
Hands	42	22	Wrist, finger tips, finger MCP

Given the set of points P, which encapsulate the x, y, z coordinates of all detected landmarks, two subsets were defined from the selected landmarks: P'_{body} and P'_{hand}. The first includes the keypoints corresponding to the nose, shoulders, and wrists, while the second contains those related to the fingertips and knuckles. From the points in both subsets, Euclidean distances were computed. Initially, non-manual features were obtained as the distances between keypoints in P'_{body}. Then, manual features were derived as the pairwise distances among the keypoints in P'_{hand}, including both intra-hand and inter-hand relations. The distance is calculated as the $L2$-norm of $p_i - p_j$, where $p_i = (X_i, Y_i, Z_i)$ and $p_j = (X_j, Y_j, Z_j)$ denote the landmark coordinates. This process results in a unique representative vector of the signer pose of 37 elements. A visual example of the computed distances in P'_{body} is shown in Figure 4.

This feature extraction process does not eliminate the temporal characteristics of the data, resulting in a sequence of vector of length n_f, being n_f equal to the number of frames of the original video. It is important to note that the number of frames varies greatly for each video, even when only accounting for an individual sign, due to differences in each signer's performance pace and in the duration of individual signs. Consequently, the features must be standardized to a fixed-length vector representation prior to being input into the model. Two different approaches were explored to achieve this normalization.

Both approaches follow the same methodology, differing only in how the value of s_l is determined. In both cases, clips with fewer frames ($n_f < s_l$) are padded with zeros until reaching the fixed duration, whereas longer clips are divided into s_l equal segments, from which a random frame is selected.

In the first approach, referred to as *average sampling*, s_l is set to the average duration of all clips. In the second approach, known as *max-length padding*, s_l is defined as the n_f of the longest video. In the latter case, no information is lost, as no frames are discarded during the sampling process. These strategies

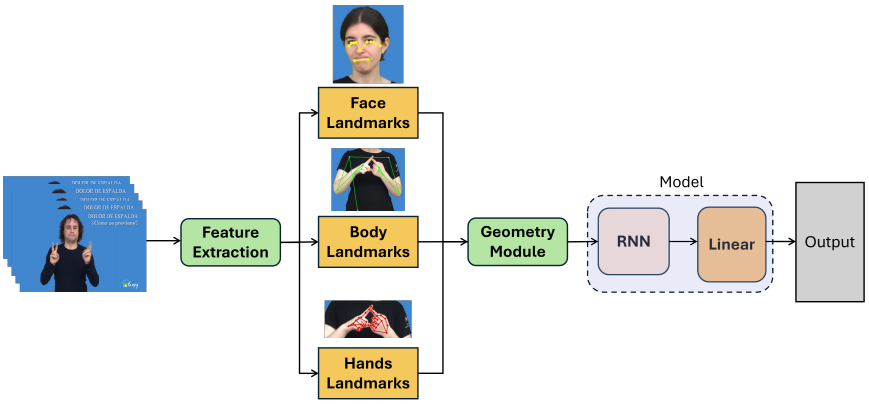

Figure 3: Diagram of the proposed pipeline for ISLR.

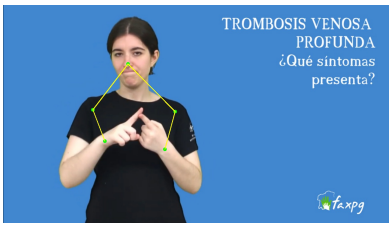

Figure 4: Connections between pose and hands Landmarks.

allowed the models to process batches of uniform size while preserving the temporal dynamics of the original gestures and ensuring consistency across samples.

4.2 Proposed Architecture

In line with previous studies on ISLR (Samaan et al., 2022; Sharma et al., 2023; Rastgoo et al., 2020), the proposed model employs Recurrent Neural Networks (RNNs) (Rumelhart and McClelland, 1987) to learn the temporal dynamics inherent in sign language. Two architectural variants were evaluated as temporal pooling mechanisms: a Bi-LSTM (Hochreiter and Schmidhuber, 1997) and a Bi-GRU (Cho et al., 2014). Both architectures process temporal sequences in forward and backward directions, allowing the model to incorporate contextual information from the entire sequence.

Figure 3 illustrates the proposed architecture, where the landmarks from three different body regions are processed to predict the sign corresponding to each input sequence. As previously described, the keypoints are converted into a single feature matrix through a series of distance calculations, performed within the *Geometry Module*. Since this module preserves the temporal characteristics of the data, a temporal pooling mechanism is required to model the temporal dependencies across frames. To this end, the model employs the previously mentioned RNN-based architectures to transform the sequential information into a compact representation. Finally, a fully connected layer aggregates the temporal features extracted by the recurrent network and projects them into a lower-dimensional space corresponding to the number of signs.

4.3 Experimental Setup

The dataset was randomly divided into three subsets following a 70–15–15 split ratio for training, validation, and testing, respectively. The partition was performed at the sample level to maintain class representation across all subsets and to avoid data leakage between training, and evaluation phases. Each video was preprocessed using the approaches described in the subsection 4.1, resulting in normalized sequences of distances between landmark-based features.

For training, two loss functions were considered: Cross-Entropy (CE) (equation 1) and Focal Loss (FL) (equation 2) (Lin et al., 2018). In both formulas, the term y_i denotes the ground-truth label for class i, taking the value 1 if the sample belongs to that class and 0 otherwise, while $\hat{y}_i$ represents the predicted probability that the sample belongs to class i. In this context, k is the total number of signs in the dataset and α_i the weight assigned to each class. The cross-entropy loss served as the baseline, while focal loss was used to address the strong class imbalance observed in the dataset. The Focal Loss introduces a modulation factor γ that reduces the influence of easily classified samples and focuses learning on more difficult or underrepresented classes. This parameter was tested with values between 2 and 3.

$$\mathcal{L}_{\text{CE}} = \sum_{i=1}^{k} -\alpha_i y_i \log(\hat{y}_i) \tag{1}$$

$$\mathcal{L}_{\text{FL}} = \sum_{i=1}^{k} -\alpha_i (1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) \tag{2}$$

To further address the class imbalance, a weighting parameter was introduced into the loss function. This parameter α_i was computed as shown in Equation 3, where n_i represents the number of samples belonging to gesture i. The class weights were applied to both loss functions by multiplying the loss value with the corresponding α_i, thereby increasing the contribution of minority classes during model optimization.

$$\alpha_i = \frac{\sum_{i=1}^{k} n_i}{k * n_i} \tag{3}$$

The models were implemented in Python using PyTorch as framework. Training was performed on a workstation (256GB of RAM), and four NVIDIA GeForce GTX 1080Ti 11GB GPUs.

4.4 Metrics

The metrics used to evaluate the performance of the models were Unweighted Accuracy (UA), F1-Score and Unweighted Average Recall (UAR). The UAR is used to measure the model capacity to correctly identify all sign classes independently of their relative frequency within the dataset, and is given by Equation 5. UA is given by the Equation 4, being the relation of the total number of true positives versus the total number of samples, and is equivalent to the commonly-named accuracy. Additionally, the F1-score was reported, defined in Equation 6, which represents the harmonic mean between precision and recall. By employing these three distinct metrics, the impact of class imbalance can be more accurately evaluated, ensuring that all classes are given comparable importance and resulting in a more balanced and representative assessment of the model's performance.

$$UA = \frac{\text{\# of successful predictions}}{\text{\# samples}} \tag{4}$$

$$UAR = \frac{1}{k} \sum_{i=1}^{k} Recall_i \tag{5}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

5 RESULTS AND DISCUSSION

Tables 2 and 3 compare Bi-LSTM and Bi-GRU performance under the two proposed pre-processing strategies: *max-length padding* and *average sampling*. *Max-length padding* achieved a 75.8% mean UA, outperforming *average sampling* (74.8%) by 1%, with similar leads in F1-score (1.1%) and UAR (0.8%). Consequently, *average sampling* can be considered the best strategy for this model. Unlike truncation methods, padding prevents information loss, allowing the model to leverage all available temporal data for more accurate predictions.

Although the *max-length padding* approach yielded slightly higher results than *average sampling approach*, suggesting that maintaining the full temporal evolution of gestures benefits recognition accuracy, it is important to note that the *average sampling* strategy remains a computationally efficient, and could be especially useful for large-scale datasets.

Two additional observations can be drawn from the results. First, the Bi-LSTM architecture consistently outperforms the Bi-GRU across all evaluated metrics, indicating its superior ability to capture temporal dependencies within the signing sequences. However, this trend is observed only when using the Focal Loss function. When trained with CE, the Bi-GRU achieves higher performance than the Bi-LSTM, likely due to its smaller number of parameters, which makes it easier to optimize under a simpler loss formulation. Second, training with the FL function yields better results than with the standard CE, emphasizing the advantages of mitigating class imbalance by reducing the influence of well-classified samples during training.

Focusing on the experiments based on the *max-length padding* strategy, the Bi-LSTM model combined with the FL function and a focusing parameter of $\gamma = 3$ achieved the best performance, reaching 79.8% UA, 79.1% F1-score, and 80.2% UAR. This configuration outperformed all other variants, indicating that an intermediate focusing parameter balances the influence of majority and minority classes. Higher values of γ led to a slight decrease in performance, suggesting excessive emphasis on difficult samples that may hinder generalization.

For the *average sampling* method, the Bi-LSTM model with FL and $\gamma = 2$ obtained 77.8% UA, 76.0% F1-score, and 77.7% UAR. Although slightly lower than those achieved with *max-length padding*, these results confirm the robustness of the architecture when handling variable-length sequences normalized through temporal resampling. As with the *max-length padding* experiments, increasing γ beyond moderate values resulted in a reduction of overall performance.

Overall, these findings validate the suitability of recurrent neural networks trained on pose-based features for ISRL. The results also emphasize the potential of this approach as a lightweight and interpretable solution for inclusive human–machine interfaces, where the goal is not only accuracy but also efficiency, and accessibility.

Table 2: Model evaluation with the max-length padding set. The best results are in bold.

Max-length Padding

Model	Loss Function	UA (%)	F1-score (%)	UAR (%)
Bi-GRU	CE	76,8	74,6	75,3
Bi-GRU	FL ($\gamma = 2$)	75,7	73,9	75,6
Bi-GRU	FL ($\gamma = 3$)	73,4	71,5	73,3
Bi-LSTM	CE	75,5	73,5	73,3
Bi-LSTM	FL ($\gamma = 2$)	73,4	73.0	73,7
Bi-LSTM	FL ($\gamma = 3$)	**79,8**	**79,1**	**80,2**
Means		75.8	74.3	75.3

Table 3: Model evaluation with the average sampling set. The best results are in bold.

Average Sampling

Model	Loss Function	UA (%)	F1-score (%)	UAR (%)
Bi-GRU	CE	76,8	74,6	75,3
Bi-GRU	FL ($\gamma = 2$)	73,3	71,5	72,8
Bi-GRU	FL ($\gamma = 3$)	72,3	70,9	72,9
Bi-LSTM	CE	75,6	73.0	74,2
Bi-LSTM	FL ($\gamma = 2$)	**77,8**	**77.0**	**77,7**
Bi-LSTM	FL ($\gamma = 3$)	72,9	71,9	73,8
Means		74.8	73,2	74.5

5.1 Model Efficiency

Given the mentioned relevance of latency in HMI scenarios, developing a lightweight model was a central priority in this work. The proposed architectures was designed to minimize computational complexity while maintaining competitive recognition performance. With a total of only 9,619 trainable parameters, the best-performing model, the Bi-LSTM, achieved an average inference time of approximately 1 ms per sample, making it well suited for real-time applications. This compact design not only reduces the computational and memory requirements but also ensures seamless and responsive interaction in user-facing systems.

6 CONCLUSION

This work presented a lightweight pose-based approach for ISLR in LSE, leveraging recurrent neural networks to model the temporal dynamics of signing sequences. The proposed pipeline combines multi-part landmark extraction using the MediaPipe framework, spatial normalization, and temporal alignment through padding and sampling strategies, resulting in standardized feature sequences suitable for sequence learning.

Among all evaluated configurations, the Bi-LSTM architecture trained with the FL function ($\gamma = 3$) and the *max-length padding* achieved the best overall performance, reaching 79.8% UA. This combination demonstrated a superior capability to capture long-term temporal dependencies while effectively mitigating class imbalance. Furthermore, the *max-length padding* strategy slightly outperformed the *average sampling* approach, likely due to the preservation of all temporal information across sequences.

The results confirm the feasibility and robustness of pose-based bidirectional recurrent architectures for ISLR, highlighting their potential to serve as lightweight and interpretable alternatives to end-to-end video-based solutions. These findings provide a solid foundation for future research in multimodal sign language understanding and gesture-driven HMI.

ACKNOWLEDGMENTS

This publication is part of the project PID2021-122402OB-C22, funded by MCIN/AEI/10.13039/501100011033/FEDER, EU, the ACIISI-Gobierno de Canarias and FEDER under project ULPGC Facilities Net and Grant EIS 2021 04, and by the Consejería de Universidades, Ciencia e Innovación y Cultura (Gobierno de Canarias) and the European Social Fund Plus (FSE+) under the funding framework for doctoral research.

REFERENCES

Al Abdullah, B. A., Amoudi, G. A., and Alghamdi, H. S. (2024). Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Access*, 12:128871–128895.

Alba-Castro, J. L., Vázquez-Enríquez, M., Pérez-Pérez, A., Mariño-Pérez, F., Lema-Álvarez, M. L., Cabeza-

Pereiro, C., Rodríguez-Banga, E., Docío-Fernández, L., Torres-Guijarro, S., Caderno-Fernández, A., and Cid-Álvarez, S. (2023). Lse-health-uvigo.

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., and Zisserman, A. (2021). Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues.

Alyami, S., Luqman, H., and Hammoudeh, M. (2024). Reviewing 25 years of continuous sign language recognition research: Advances, challenges, and prospects. *Information Processing Management*, 61(5):103774.

Attig, C., Rauh, N., Franke, T., and Krems, J. F. (2017). System latency guidelines then and now – is zero latency really considered necessary? In *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, Cham. Springer International Publishing.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2021). Multi-channel transformers for multi-articulatory sign language translation. volume 12538 of *Lecture Notes in Computer Science*. Springer International Publishing.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Cooper, H., Ong, E.-J., Pugeault, N., and Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13(72):2205–2231.

Coster, M. D., Herreweghe, M. V., and Dambre, J. (2021). Isolated sign recognition from rgb video using pose flow and self-attention. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3436–3445.

Dautenhahn, K. (1995). Getting to know each other—artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16(2):333–356. Moving the Frontiers between Robotics and Biology.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i Nieto, X. (2021). How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, D., Yi, M., Kang, W., Wang, Y., and Lv, C. (2024). Continuous sign language recognition algorithm based on object detection and variable-length coding sequence. *Sci. Rep.*, 14(1):27592.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Koller, O., Forster, J., and Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141.

Laines, D., Bejarano, G., González-Mendoza, M., and Ochoa-Ruiz, G. (2023). Isolated sign language recognition based on tree structure skeleton images. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Li, D., Opazo, C. R., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines.

Lv, Z., Poiesi, F., Dong, Q., Lloret, J., and Song, H. (2022). Deep learning for intelligent human–computer interaction. *Applied Sciences*, 12(22).

Niu, Z. and Mak, B. (2020). Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *Computer Vision – ECCV 2020*, Cham. Springer International Publishing.

Rastgoo, R., Kiani, K., and Escalera, S. (2020). Video-based isolated hand sign language recognition using a deep cascaded model. *Multimed. Tools Appl.*, 79(31-32):22965–22987.

Rumelhart, D. E. and McClelland, J. L. (1987). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA.

Samaan, G. H., Wadie, A. R., Attia, A. K., Asaad, A. M., Kamel, A. E., Slim, S. O., Abdallah, M. S., and Cho, Y.-I. (2022). MediaPipe's landmarks with RNN for dynamic sign language recognition. *Electronics (Basel)*, 11(19):3228.

Sarhan, N. and Frintrop, S. (2023). Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

Saunders, B., Camgoz, N. C., and Bowden, R. (2021). Continuous 3D Multi-Channel sign language production via progressive transformers and mixture density networks. *International Journal of Computer Vision*, 129(7):2113–2135.

Sharma, S., Gupta, R., and Kumar, A. (2023). Continuous sign language recognition using isolated signs data and deep transfer learning. *J. Ambient Intell. Humaniz. Comput.*, 14(3):1531–1542.

Sutton-Spence, R. and Woll, B. (1999). *The linguistics of British sign language: An introduction*. Cambridge University Press.

Tarrés, L., Gállego, G. I., Duarte, A., Torres, J., and Giró-i Nieto, X. (2023). Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5625–5635.

Wang, Z., Li, D., Jiang, R., and Okumura, M. (2025). Continuous sign language recognition with multi-scale spatial-temporal feature enhancement. *IEEE Access*.

Yuan, T., Sah, S., Ananthanarayana, T., Zhang, C., Bhat, A., Gandhi, S., and Ptucha, R. (2019). Large scale sign language interpretation. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–5.