Which Feature-Classifier Combinations for Video Sign Language Recognition?

Luan Fernandes De Franca[1], José Everardo Bessa Maia[1]

LADESC - Universidade Estadual do Ceará, Brazil
luan.franca@aluno.uece.br

Abstract. This paper presents a comparative investigation of five feature extractors frequently used for video-based sign language recognition: I3D-RGB, I3D-OF, MediaPipe, Pose, TSM, and SlowFast. Many studies utilize these features, yet there is a lack of comparative analysis regarding their complexity and discriminative power when combined with linear, non-linear, and deep neural classifiers. Test results based on the public UFOP-LIBRAS-ISO dataset indicate that the top two performances were achieved by the TMS+LR combination (F1-score of 99.30%) and the SlowFast+LR combination (F1-score of 98.40%). Conversely, the complexity ranking begins with Pose as the least computationally demanding, followed by TSM, I3D-RGB, I3D-OF, MediaPipe, and SlowFast, with the latter being the most complex.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: LIBRAS, Sign Language Recognition, Transfer Learning, Deep Learning, Feature Extraction

1. INTRODUCTION

Video-based sign language recognition is a significant application task, as evidenced by the volume of papers published in recent years [Singh et al. 2025; Sarhan and Frintrop 2020; Boháček and Hrúz 2022]. The recognition task is framed as a classification problem involving one of several predefined classes. In video-based sign language recognition, this task is characterized by high dimensionality, both in the input space and in the large number of output classes.

Due to the high dimensionality of video data, the classification step is invariably preceded by a feature extraction step. Following a period dominated by hand-engineered features, our survey indicates that recent literature focuses on five techniques for extracting video features for sign language recognition: I3D-RGB and I3D-OF [Singh et al. 2025; Sarhan and Frintrop 2020], Pose (skeleton) [Boháček and Hrúz 2022], SlowFast [Ahn et al. 2024], and TSM [Zhu et al. 2022]. With the exception of skeleton-based (Pose) features, these techniques rely on features derived from the internal layers of deep neural networks pre-trained on large volumes of generic data; will be described in greater detail in a subsequent section.

However, the literature lacks a fair performance comparison regarding the discriminative capacity of these feature extractors when combined with linear, non-linear, and deep neural classifiers—a comparison that would facilitate the development of selection guidelines. In this work, we evaluate the performance of various combinations of these feature extractors and classifiers using the UFOP-LIBRAS-ISO dataset, aiming to rank the combinations based on their performance.

The remainder of this article is organized as follows. Section 2 briefly reviews related work, while Section 3 outlines the fundamentals of the adopted feature extractors and classifiers. Section 4 presents the experimental design and results. Section 5 concludes the article.

2. RELATED WORK

[Cerna et al. 2021] introduced the **LIBRAS-UFOP** dataset, a publicly available and challenging resource for Brazilian Sign Language (Libras) recognition, collected using a Microsoft Kinect V1 sensor. The dataset comprises 56 signs organized into four categories based on the concept of minimal pairs signs with high similarity that differ in only one primary parameter (hand configuration, articulation point, or movement) totaling 3,040 RGB-D and skeleton sequences performed by five participants. For validation, the authors proposed a multimodal baseline method that generates dynamic images from RGB-D and skeleton data, used as input to two multi-stream CNN architectures (3S-RGBD-CNN and 3S-SKL-CNN). Results obtained with late fusion reached a recognition rate of 74.25%, outperforming traditional methods such as SC-HCM (63.30%) and pose-based methods such as P-CNN (68.14%), highlighting the challenging nature of the dataset due to the high similarity between signs.

[Renjith et al. 2026] proposed TransMoVIF, a dual-stream transformer-based architecture for isolated sign language recognition that combines skeletal motion trajectories and visual appearance cues via cross-attention. A crucial frame selection mechanism, combining velocity-based analysis and the Structural Similarity Index (SSIM), reduces redundant frames by approximately 87.5% while preserving discriminative gesture transitions. The pose trajectory encoder captures displacement, velocity, and curvature from hand keypoints, while a CNN-based visual encoder processes the selected RGB frames. Unlike conventional late-fusion approaches, the cross-attention mechanism enables deep bidirectional interaction between the motion and appearance streams. The authors evaluated the model on the Include-50 (Indian Sign Language) and SMILE-DSGS (Swiss German Sign Language) datasets, achieving accuracies of 96.80% and 93.95%, respectively, outperforming prior state-of-the-art methods such as CNN+LSTM and ST-GCN-based approaches.

[Desai et al. 2023] introduced **ASL Citizen**, the first crowdsourced Isolated Sign Language Recognition (ISLR) dataset, containing 83,399 videos of 2,731 distinct signs performed by 52 signers in everyday, uncontrolled environments. Unlike prior datasets collected in laboratory settings or scraped from the web, ASL Citizen was gathered with informed consent and direct involvement of the Deaf community, targeting the task of sign language dictionary retrieval, in which users demonstrate a sign via webcam to retrieve matching entries. The authors trained appearance-based (I3D) and pose-based (ST-GCN) classifiers, evaluated on completely unseen signers, achieving 63.10% top-1 accuracy and 90.86% recall-at-10 with the I3D model more than double the accuracy of models trained on the previously largest dataset, WLASL-2000, under comparable evaluation conditions.

[Li et al. 2020] presented the **WLASL** (Word-Level American Sign Language) dataset, the largest public ASL dataset available at the time, comprising 21,083 videos of 2,000 glosses performed by 119 signers, collected from educational websites and YouTube tutorials. The dataset is organized into four subsets of increasing vocabulary size (WLASL100, WLASL300, WLASL1000, and WLASL2000) to study the scalability of recognition methods. The authors benchmarked appearance-based (VGG-GRU, I3D) and pose-based (Pose-GRU) baselines, and additionally proposed **Pose-TGCN**, a temporal graph convolutional network that jointly models spatial and temporal dependencies among body keypoints. The fine-tuned I3D model achieved the best performance, reaching 32.48% top-1 accuracy and 66.31% top-10 accuracy on the full WLASL2000 vocabulary, while Pose-TGCN achieved comparable results at top-5 and top-10 accuracy despite its significantly smaller model size.

3. THEORETICAL BACKGROUND

3.1 Classifiers

Sign language classification is addressed through both classical Machine Learning and Deep Learning approaches, including: Logistic Regression (LR), Convolutional Neural Network (CNN) [LeCun et al. 2015], and Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber 1997]. The

`scikit-learn` library [Pedregosa et al. 2011] was used to implement Logistic Regression, while CNN and LSTM models were developed using the PyTorch library [Paszke et al. 2019].

3.2 Classification Metrics

Classification metrics quantify the performance of models that assign discrete labels to input instances, derived from the four fundamental values of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [Fawcett 2006]. **Precision** ($TP/(TP + FP)$) measures the proportion of correct positive predictions; **Recall** ($TP/(TP+FN)$) measures the model's ability to correctly identify all instances of the positive class; and the **F1-Score** combines both through their harmonic mean, being particularly useful in class-imbalanced scenarios [van Rijsbergen 1979].

Finally, **AUC-ROC** [Fawcett 2006] quantifies the model's discriminative ability across all possible decision thresholds, ranging from 0.5 (random classifier) to 1.0 (perfect discrimination). Due to space limitations, the reader is referred to the cited references for further formal details on these metrics.

3.3 Pre-Trained Features

The **I3D** (*Inflated 3D ConvNet*) architecture, proposed by [Carreira and Zisserman 2018], extends 2D convolutional networks to the temporal domain by inflating two-dimensional filters into three-dimensional ones. This approach enables the simultaneous modeling of spatial and temporal information from video sequences, making it particularly effective for video analysis. The I3D model was originally trained on the Kinetics dataset for action recognition tasks. For feature extraction, the output of the last convolutional layer before the final classification layer is used, producing a 1,024-dimensional feature vector that encodes spatiotemporal information extracted from the video frames. I3D has variants for RGB input (3 channels) and optical flow (2 channels), both generating feature vectors of the same dimensionality.

The **SlowFast** network, proposed by [Feichtenhofer et al. 2019], is a dual-pathway architecture designed for video understanding. The Slow pathway operates at a low frame rate, capturing spatial semantics, while the Fast pathway operates at a high frame rate, capturing motion and temporal dynamics with reduced channel capacity. The two pathways are fused through lateral connections, combining complementary spatiotemporal representations. For feature extraction, the output of the last pooling layer is used, generating a feature vector that encodes both semantic and motion information from the video frames. SlowFast has demonstrated strong performance on action recognition benchmarks such as Kinetics and AVA.

The **TSM** (*Temporal Shift Module*), proposed by [Lin et al. 2019], is a lightweight and efficient architecture for video understanding that introduces temporal reasoning into 2D convolutional networks without adding computational overhead. The core idea consists of shifting part of the feature map channels along the temporal dimension, allowing each frame to exchange information with its neighboring frames. This mechanism enables the network to capture temporal dependencies while maintaining the efficiency of standard 2D convolutions. TSM can be inserted into any existing 2D CNN backbone, such as ResNet, making it highly flexible and easy to deploy. Despite its simplicity, TSM achieves competitive performance on action recognition benchmarks such as Kinetics and Something-Something, requiring significantly less computation than 3D convolutional approaches such as I3D and SlowFast.

3.4 Skeleton Features

MediaPipe Holistic [Lugaresi et al. 2019] is an open-source framework developed by Google that provides integrated and simultaneous tracking of the body, hands, and face from 2D video. The Holistic module combines three submodules into a single inference pipeline — Pose, Hands, and Face

Mesh — producing a comprehensive set of anatomical landmarks per video frame.

Pose: The pose submodule detects 33 body joint landmarks, including shoulders, elbows, wrists, hips, knees, and ankles. Each landmark is represented by coordinates $(x, y, z, \text{visibility})$, yielding a feature vector of $33 \times 4 = 132$ values per frame. The x and y coordinates are normalized by image width and height, z represents the estimated relative depth, and *visibility* encodes detection confidence.

Hands: The hand submodule tracks 21 landmarks per hand (left and right separately), covering the wrist and four articulations of each finger. Each landmark is represented by (x, y, z) coordinates, resulting in vectors of $21 \times 3 = 63$ values per hand, or 126 values for both hands combined. These keypoints capture hand configuration, finger orientation, and wrist posture primary parameters that are fundamental in sign languages.

Face Mesh: The facial mesh submodule tracks 468 landmarks distributed across the face, including eye contours, eyebrows, nose, mouth, and cheeks. Each landmark is represented by (x, y, z) coordinates, producing a vector of $468 \times 3 = 1{,}404$ values per frame. This set of landmarks enables the analysis of facial expressions and nonmanual markers, which play a relevant grammatical role in sign languages [Kuznetsova and Kimmelman 2024].

The complete Holistic feature vector per frame, considering all submodules, totals:

$$\underbrace{33 \times 4}_{\text{Pose}} + \underbrace{21 \times 3 \times 2}_{\text{Hands}} + \underbrace{468 \times 3}_{\text{Face}} = 132 + 126 + 1{,}404 = 1{,}662 \text{ values} \tag{1}$$

For a sequence of T frames, the resulting input tensor has dimension $T \times 1{,}662$, which can be reorganized and subsampled according to the requirements of the recognition task. In practice, it is common to use only subsets of this vectorfor instance, only hand and pose landmarks to reduce dimensionality and focus on the articulations most relevant to isolated sign recognition.

4. METHODOLOGY

4.1 Dataset

The **LIBRAS-UFOP-ISO** dataset comprises 56 classes of isolated Libras signs. Data were collected from five participants (three women and two men), each performing approximately ten repetitions per sign, totaling 3,040 RGB-D data sequences. Recordings were made using a Kinect device positioned 2 meters from the participants, capturing full-body videos. Available at: https://www.repositorio.ufop.br/items/07e29918-101d-4716-8c20-f297e2115918/request-a-copy

4.2 Glossary Term Recognition

From the dataset described in Section 4.1, whose `.txt` annotations were processed by the segmentation model to isolate individual signs, 16-frame video clips were generated at a resolution of 224×224 pixels, with frames uniformly sampled along each sign. Frames discarded by the segmentation process corresponding to moments without significant hand movement were used to compose an additional class termed *neutral/rest*, representing instants in which the individual performs no sign and remains in a static position, yielding a total of 57 output classes.

The processed data were submitted to five distinct feature extraction approaches, whose computational complexity is compared in Table I.

The first approach used the I3D-RGB network to extract features from the RGB stream, producing a **1,024-dimensional** feature vector per clip that encodes dense spatiotemporal information of the gestures. The second employed the same I3D architecture applied to Optical Flow (I3D-OF),

also producing **1,024-dimensional** vectors, explicitly capturing motion patterns between consecutive frames rather than visual appearance. The third approach used the TSM (*Temporal Shift Module*) network, which operates on sequences of 16 RGB frames at 224×224 pixels resolution, introducing temporal modeling through channel shifting between consecutive frames without additional computational cost relative to the ResNet-50 backbone. The resulting feature vector has **2,048 dimensions**. The fourth approach employed the SlowFast network, which processes two parallel temporal pathways one slow, focused on spatial semantics, and one fast, focused on motion dynamics producing a **2,304-dimensional** feature vector. As the architecture requires a minimum of 32 input frames, each 16-frame clip was temporally duplicated to satisfy this requirement, maintaining the 224×224 pixel.

The fifth approach employed MediaPipe Holistic for structured skeleton-based feature extraction. Unlike the four previous approaches, which use deep networks to produce dense feature vectors from raw pixels, MediaPipe Holistic operates exclusively on CPU and directly outputs anatomical landmark coordinates of the body, hands, and face totaling **1,662 values per frame**, as described in Section 3.4. This approach does not require a dedicated GPU and has a lower memory footprint, making it suitable for hardware-constrained scenarios, although it processes one frame at a time at approximately 4.8 frames per second on CPU.

Table I: Computational complexity comparison of the evaluated feature extractors. GFLOPs refer to a single input clip at 224×224 pixel resolution. Model size refers to pre-trained weights stored on disk. The feature vector dimension corresponds to the output used as input to the classifiers.

Method	**Param.**	**GFLOPs**	**Model (MB)**	**Input**	**Features**
I3D-RGB [Carreira and Zisserman 2018]	28.0 M	37.3	∼107	16×RGB	1,024
I3D-OF [Carreira and Zisserman 2018]	28.0 M	37.3	∼107	16×OF	1,024
TSM (ResNet-50) [Lin et al. 2019]	24.3 M	32.9*	∼92	16×RGB	2,048
SlowFast (R-50) [Feichtenhofer et al. 2019]	34.6 M	66.1	∼130	32×RGB**	2,304
MediaPipe Holistic [Lugaresi et al. 2019]	<10 M	<5***	∼25	1 frame (CPU)	1,662

*TSM does not add floating-point operations to the ResNet-50 backbone; GFLOPs are therefore equivalent to the base network.

**Original 16-frame clips were temporally duplicated to satisfy the minimum 32-frame input requirement of SlowFast.

***Approximate values; the pipeline combines multiple lightweight models in cascade (BlazePose, Palm Detector, and Face Mesh), whose individual costs are not publicly reported in a consolidated form.

Glossary term recognition is treated as a 57-class multiclass classification task. The combinations evaluated in this work were: TSM+LR, I3D-RGB+LR, I3D-RGB+CNN, SlowFast+LR, SlowFast+CNN, Pose+LSTM, I3D-OF+LR, and I3D-OF+CNN. The computational complexity of each classifier, estimated for their respective input dimensions, is presented in Table II. The adopted hyper parameters are detailed in Table III. For Logistic Regression, L2 regularization with up to 1,500 iterations was adopted to ensure convergence in high-dimensional spaces. The 1D CNN was configured with three convolutional layers of 256 channels and a dropout of 0.3 to mitigate overfitting. The LSTM was defined with two recurrent layers of hidden size 256 and a dropout of 0.3, aiming to capture temporal dependencies along the feature sequence. All deep models were trained using the Adam optimizer.

Table II: Computational complexity of the evaluated classifiers. Parameters estimated for 57 classes and input dimension according to the extractor: 1,024 (I3D-RGB, I3D-OF), 1,662 (Pose/MediaPipe), and 2,048/2,304 (TSM/SlowFast).

Classifier	**Parameters**	**Complexity (train)**	**Complexity (inf.)**	**Optimizer**
LR [Pedregosa et al. 2011]	∼58 K–133 K	$O(n \cdot d)$	$O(d)$	L-BFGS
CNN [LeCun et al. 2015]	∼1.2 M	$O(T \cdot K \cdot C^2)$	$O(T \cdot K \cdot C)$	Adam
LSTM [Hochreiter and Schmidhuber 1997]	∼2.1 M	$O(T \cdot d \cdot h)$	$O(T \cdot h)$	Adam

n: training samples; d: input vector dimension; T: number of frames; K: kernel size; C: channels; h: hidden state dimension. The parameter range for LR reflects the varying input dimensions across the evaluated extractors.

5. RESULTS

Table IV presents the isolated sign classification results for the different combinations of feature extractors and classifiers evaluated. In all direct comparisons using the same feature base, Logistic

Table III: Hyperparameters of the classifiers used in the experiments. Logistic Regression was implemented with `scikit-learn` [Pedregosa et al. 2011], while CNN and LSTM were developed using PyTorch [Paszke et al. 2019].

Method	**Hyperparameters**
LR	`max_iter=1500; penalty='l2'`
CNN	`channels=256; kernel_size=3; num_layers=3; dropout=0.3`
LSTM	`hidden_size=256; num_layers=2; dropout=0.3`

Regression (LR) outperformed more complex classifiers. The best overall results were achieved by TSM+LR (F1-score = 0.993), followed by SlowFast+LR (F1-score = 0.984) and I3D-RGB+LR (F1-score = 0.977). In contrast, CNN-based variants using the same feature base performed worse: I3D-RGB+CNN achieved an F1-score of 0.947 and SlowFast+CNN of 0.866. A likely explanation for this pattern is that the pre-trained deep networks (I3D, SlowFast, TSM) already produce linearly separable representations, making a linear classifier such as LR sufficient to exploit this structure. Classifiers with a larger number of trainable parameters, such as CNN and LSTM, tend to overfit given the limited data volume available in the dataset used.

The Pose+LSTM model achieved the highest AUC-ROC of 1.000 among all methods, yet with an F1-score of only 0.750 and high variance across runs (standard deviation of 6.75). This behavior indicates that, although the model discriminates well between classes in terms of ranking (high AUC), its absolute classification accuracy is lower than that of appearance-based extractors, and its instability suggests sensitivity to random data partitioning. These results indicate that skeleton-based features alone do not compete with deep appearance representations in the evaluated context.

The results also show that features extracted from the RGB stream consistently outperform those extracted from Optical Flow (OF) within the same I3D architecture. The I3D-RGB+LR combination achieved an F1-score of 0.977, while I3D-OF+LR obtained only 0.768 — a drop of approximately 21 percentage points. This pattern is repeated with the CNN classifier (I3D-RGB: 0.947 vs. I3D-OF: 0.756). These results suggest that, for isolated Libras signs, visual appearance — hand configuration, body posture, and gesture location — is more discriminative than the pure motion captured by optical flow, thus providing a negative answer to the hypothesis of OF complementarity to RGB features.

Método	AUC-ROC	Precision	Recall	F1-score
I3D-RGB + LR	0.999±0,0000 (0.999;0.999)	0.978±0.0035 (0.970;0.983)	0.977±0.0035 (0.969;0.981)	0.977±0.0037 (0.969;0.993)
I3D-RGB + CNN	0.999±0.0001 (0.999;0.999)	0.956±0.0061 (0.946;0.966)	0.949±0.0066 (0.935;0.958)	0.947±0.0069 (0.933;0.981)
I3D-OF + LR	0.992±0.0003 (0.991;0.992)	0.779±0.0076 (0.766;0.793)	0.767±0.0089 (0.749;0.782)	0.7676±0.0087 (0.750;0.780)
I3D-OF + CNN	0.993±0.0005 (0.992;0.993)	0.775±0.0076 (0.759;0.786)	0.759±0.0081 (0.745;0.769)	0.756±0.0082 (0.745;0.767)
Pose + LSTM	1.000±0.0000 (0.999;1.000)	0.767±6.89 (0.668;0.876)	0.786±5.36 (0.721;0.864)	0.750±6.75 (0.667;0.851)
SlowFast + CNN	0.939±0.020 (0.888;0.958)	0.879±0.050 (0.754;0.926)	0.881±0.040 (0.781;0.919)	0.866±0.045 (0.756;0.910)
SlowFast + LR	0.999±0.0000 (0.999;1.000)	0.985±0.003 (0.978; 0.989)	0.984±0.003 (0.977;0.988)	0.984±0.003 (0.977;0.988)
TSM + LR	1.000±0.0000 (0.999;1.000)	0.994±0.002 (0.989;0.998)	0.993± 0.002 (0.988;0.998)	0.993± 0.002 (0.988;0.998)

Table IV: Classification performance for isolated sign language videos. Each video was initially sampled by extracting 16 equally spaced frames, generating a reduced clip to which the methods listed in the table were applied. Statistical data are reported over 10 runs with random data subsets: 80% for training and 20% for testing.

To validate the models in a scenario closer to real-world use, the models with the highest F1-scores were evaluated on a continuous 30-second video in which an interpreter performs the same sign ten consecutive times, separated by neutral pose intervals. The video was segmented into clips and submitted to the models for classification, whose results are presented in Figure 1.

The TSM and SlowFast models, despite achieving the highest F1-scores in the isolated classification experiments, performed below expectations in the continuous video test. The TSM model incorrectly classified all clips corresponding to the rest class, assigning them to other classes in the dataset. The SlowFast model, in turn, demonstrated better performance on the rest class, but exhibited a high error rate on the target sign class, classifying signs as other classes rather than the correct one.

The I3D-RGB+LR model achieved the best performance in this scenario, correctly labeling ap-

proximately 95% of clips. The observed errors were concentrated at the boundary between the target class and the rest class, which can be explained by the presence of transition frames between the sign and the neutral pose, in which both actions coexist within the same clip, making classification more difficult. This behavior makes I3D-RGB+LR more suitable for continuous video applications than the models with higher F1-scores on isolated signs.

Notably, the Pose+LSTM model, which achieved the lowest F1-score in the isolated classification experiments, reached an accuracy above 98% in the continuous video test. The observed errors were concentrated in transition clips between sign and neutral pose, in which the overlap of frames from both actions introduces classification ambiguity — analogous behavior to that observed in the I3D-RGB+LR model.

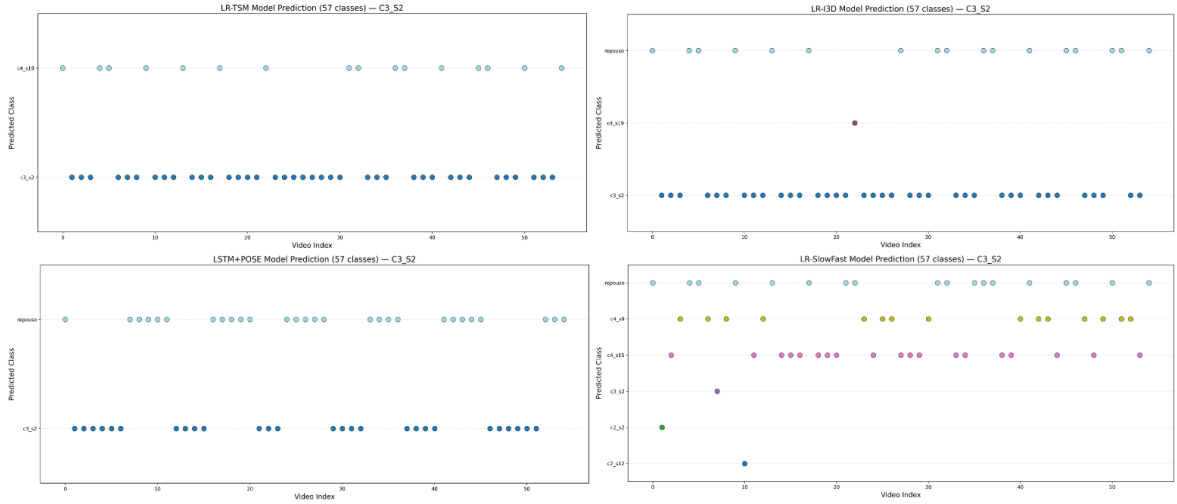

Fig. 1: Model predictions on a continuous video containing an isolated sign repeated 10 times, separated by neutral pose intervals.

6. CONCLUSION

In summary, the best-performing model in isolated classification was **TSM+LR**, with an F1-score of 0.993 and AUC-ROC of 1.000, combining the temporal efficiency of TSM with the stability and generalization of Logistic Regression. These results indicate that, for isolated Libras sign recognition with limited data, lightweight temporal feature extractors combined with linear classifiers constitute the most effective strategy.

However, performance in isolated classification does not necessarily predict model behavior in continuous scenarios. In the continuous video validation, TSM+LR did not prove to be the best approach: despite its feature extraction being more lightweight than I3D-RGB and SlowFast, and its training performance being superior, the model failed in this scenario by incorrectly classifying rest class clips. The SlowFast model exhibited similar behavior, failing to correctly label most clips of the target sign class. The I3D-RGB+LR model demonstrated the best performance in continuous video, correctly labeling approximately 95% of clips, with errors concentrated in transition frames between sign and neutral pose situations in which both actions coexist within the same clip, introducing natural classification ambiguity. This result highlights that robustness to the rest class is a critical factor for practical applications in continuous scenarios.

The Pose+LSTM model, despite achieving the lowest F1-score in the isolated classification experiments, reached an accuracy above 98% in the continuous video test, with errors concentrated in transition clips between sign and neutral pose. This result is particularly promising for real-time applications on mobile devices, given that MediaPipe Holistic the basis of the skeleton feature extraction

operates exclusively on CPU and provides a JavaScript API, enabling its direct integration into web and mobile applications without the need for a dedicated GPU.

As future work, we intend to integrate this recognition layer into a complete architecture for translating Libras into natural language via LLMs, investigate the system's operation in real time, and explore Vision Transformer (ViT)-based architectures for feature extraction, which have shown promising results in action and gesture recognition tasks in recent literature.

REFERENCES

Ahn, J., Jang, Y., and Chung, J. S. Slowfast network for continuous sign language recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3920–3924, 2024.

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

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. https://arxiv.org/abs/1705.07750, 2018.

Cerna, L. R., Cardenas, E. E., Miranda, D. G., Menotti, D., and Camara-Chavez, G. A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications* vol. 167, pp. 114179, 2021.

Desai, A., Berger, L., Minakov, F. O., Milan, V., Singh, C., Pumphrey, K., Ladner, R. E., Daumé III, H., Lu, A. X., Caselli, N., and Bragg, D. ASL citizen: A community-sourced dataset for advancing isolated sign language recognition. In *arXiv preprint*, 2023.

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8): 861–874, 2006.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 6202–6211, 2019.

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

Kuznetsova, A. and Kimmelman, V. Testing MediaPipe holistic for linguistic analysis of nonmanual markers in sign languages, 2024.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature* 521 (7553): 436–444, 2015.

Li, D., Rodriguez Opazo, C., Yu, X., and Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *arXiv preprint arXiv:1910.11006*, 2020.

Lin, J., Gan, C., and Han, S. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

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. MediaPipe: A framework for building perception pipelines, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. Curran Associates, Inc., pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* vol. 12, pp. 2825–2830, 2011.

Renjith, S., Varghese, A., Rashmi, M., and Poorna, S. Transformer-based motion-visual integrated fusion for isolated sign language recognition. *Computers and Electrical Engineering* vol. 130, pp. 110902, 2026.

Sarhan, N. and Frintrop, S. Transfer learning for videos: From action recognition to sign language recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 1811–1815, 2020.

Singh, S., Keserwani, P., Inoue, K., Iwamura, M., and Roy, P. P. Revisiting i3d for sign language recognition. *IEICE Transactions on Information and Systems*, 2025.

van Rijsbergen, C. J. Information retrieval. *Butterworth*, 1979.

Zhu, Q., Li, J., Yuan, F., and Gan, Q. Temporal superimposed crossover module for effective continuous sign language. *arXiv preprint arXiv:2211.03387*, 2022.