Automatic Time-aware Recognition of Brazilian Sign Language Based on Dynamic Time Warping

Lucas de S. Arcanjo
larcanjo@sga.pucminas.br
Laboratory of Image and Multimedia Data Science (IMScience),
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais

Lucas F. Coelho
lucas.coelho.1296135@sga.pucminas.br
Laboratory of Image and Multimedia Data Science (IMScience),
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais

Silvio Jamil F. Guimarães
sjamil@pucminas.br
Laboratory of Image and Multimedia Data Science (IMScience),
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais

Zenilton K. G. do Patrocínio Jr
zenilton@pucminas.br
Laboratory of Image and Multimedia Data Science (IMScience),
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais

Leonardo Vilela Cardoso
leonardocardoso@pucminas.br
Laboratory of Image and Multimedia Data Science (IMScience),
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Minas Gerais

ABSTRACT

The Brazilian Sign Language (Libras) is a crucial communication medium for the deaf community in Brazil, yet it poses significant challenges for recognition and translation tasks. This paper presents a novel approach using Fast Dynamic Time Warping (FastDTW)[1] for recognizing Libras signs in video streams. This approach aims to bridge the communication gap between deaf and hearing individuals, enhancing accessibility and reducing social marginalization. The methodology leverages MediaPipe to extract key hand and body landmarks, which are then used to compute angular features for accurate sign recognition. Experiments were conducted on the MINDS-Libras dataset, and the results demonstrated a high recognition accuracy, outperforming traditional methods. Furthermore, when the proposed model is applied to the INCLUDE-50 dataset containing signs from a different sign language, it performs competitively without relying on deep learning techniques.

KEYWORDS

Computer Vision, Sign Language Recognition, Gesture Recognition, Dynamic Time Warping, MediaPipe, Libras, Brazilian Sign Language.

1 INTRODUCTION

The Brazilian Sign Language (Libras[2]) recognition is a complex research field that has attracted considerable interest in computer vision and multimedia communities [2, 21]. One of the challenges in Libras recognition task is the content description of signers based on ground truth (GT) annotations created by multiple individuals. The variability introduced by multiple annotators often results in a GT containing diverse perspectives of the events depicted in the signer's video, thereby highlighting different body movements according to the annotators' fluency in Libras [9, 22]. According to the Brazilian Institute of Geography and Statistics – IBGE [19], about 10 million people have hearing problems, with approximately 3 million being completely deaf and living in Brazil. While communication applications and tools have been widely developed in recent decades, deaf people face numerous problems using these technologies. Outside the technological field, communication barriers also manifest, and this is the greatest difficulty in providing services to hearing-impaired individuals [21, 25].

The communication barrier faced by deaf individuals hinders equitable access to essential services. A system utilizing pattern recognition techniques could bring significant benefits to communication between deaf and hearing people, facilitating interaction in various contexts and contributing to the reduction of the marginalization of this community.

Two different strategies can be applied in sign language recognition: device-based and computer vision-based methods [2, 21]. Device-based approaches utilize specialized hardware such as data gloves, depth-sensing cameras, and other wearable sensors to capture sign language gestures [22]. These devices can provide precise data but often at the cost of user comfort and affordability. On the other hand, computer vision-based methods leverage regular cameras or webcams to capture gestures, offering a more natural and cost-effective solution [8]. These methods often employ neural networks and diverse machine-learning techniques to recognize signs.

A key distinction across computer vision works is the use of spatio-temporal features. Some of them rely exclusively on images. For this task, Convolutional Neural Networks (CNNs) have shown high accuracy in recognizing static signs, achieving up to 99.90% accuracy on grayscale images [13, 30]. However, in video, temporal

[1]Code available on https://github.com/IMScience-PPGINF-PucMinas/libras-sign-recognition
[2]In Brazilian Portuguese – Língua Brasileira de Sinais

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

data processing is critical for sign recognition. The order of the gestures is crucial, and the static processing of an image does not solve the real communication problem.

The main idea of this study is to enhance the accuracy of recognizing Libras signs from video streams and translating them into Portuguese words using FastDTW techniques. This task involves several subtasks, including signs interpretation for hands, fingers, torso, and the positioning of phalanges within images. Consequently, this work seeks to identify specific Libras gestures within video sequences and accurately map them to their corresponding Portuguese words, thereby bridging the two languages. Additionally, a central component of this research is to evaluate the effectiveness of FastDTW in this translation process. The goal is not only to assess the technique's feasibility but also to optimize its practical application for facilitating communication between Libras users and Portuguese speakers.

This paper is structured as follows. The theoretical background is presented in Section 2, while Section 3 discusses the related works. Section 4 details the proposed method, while Section 5 presents and analyzes the results. Finally, Section 6 draws some conclusions.

2 BACKGROUND

2.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is a technique for measuring the similarity between time series, introduced by Sakoe and Chiba [23] initially for speech applications. Time series analysis is a critical task in various domains. Accurate measurement of the similarity between time series is essential for classification, clustering, and anomaly detection. Two common techniques for measuring similarity between time series are Euclidean Distance (ED) and DTW. While ED is simple and efficient, it has significant limitations that DTW addresses more effectively.

Figure 1 demonstrates the difference between ED and DTW. Comparing the two signals, it is possible to observe the distortion caused by ED, whereas DTW tends to capture the temporal relationships between the two compared series. ED is a straightforward method for calculating the similarity between time series. Given time series $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ of equal length n, the ED is defined as:

$$ED(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \quad (1)$$

While ED is computationally efficient, it is sensitive to shifts and distortions in the time axis. Thus, if one time series is a slightly shifted version of another, ED may indicate a large dissimilarity, even if the series appears visually similar.

DTW was designed to overcome the limitations of ED by allowing for elastic shifting along the time axis. This makes DTW more robust to variations in time series that are misaligned or have different lengths [12]. Given time series $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_m]$, the DTW algorithm constructs an $n \times m$ cost matrix D where each element $D(i, j)$ represents the squared distance between points x_i and y_j:

$$D(i, j) = (x_i - y_j)^2. \quad (2)$$

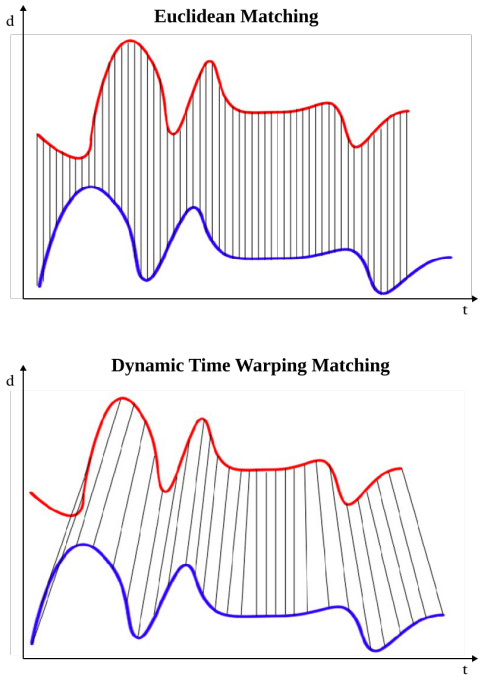

Figure 1: Comparison between DTW and ED [7], in which d represents a distance between two signs, as degree difference, and t the time variation

.

A warping path W is defined as a sequence of matrix elements representing a mapping between X and Y:

$$W = [w_1, w_2, \ldots, w_K], \quad \text{where} \quad w_k = (i, j). \quad (3)$$

The warping path must satisfy the following conditions: (i) *boundary condition*: the path starts at the bottom-left corner and ends at the top-right corner of the matrix; (ii) *continuity*: the path steps must be contiguous; and (iii) *monotonicity*: the path indices must be non-decreasing. The idea is to find the path that minimizes the cumulative distance:

$$DTW(X, Y) = \min_{W} \sqrt{\sum_{k=1}^{K} D(w_k)}. \quad (4)$$

The optimal path is determined using dynamic programming. The recursive formula to fill the cost matrix is:

$$D(i, j) = (x_i - y_j)^2 + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\}. \quad (5)$$

The FastDTW algorithm, introduced by [24], is an optimized version of the original DTW. Unlike DTW, which requires filling the entire cost matrix and has a computational complexity of $O(N^2)$, FastDTW achieves $O(N)$ complexity by strategically reducing the number of calculations needed to fill the cost matrix. This significant reduction in computational overhead makes FastDTW much faster.

The FastDTW algorithm comprises three key operations: (i) coarsening, which shrinks the time series into a smaller representation by averaging adjacent pairs of points, effectively halving the size of the series; (ii) projection, which uses the warp path from a lower

resolution as an initial guess for the higher resolution; and (iii) refinement, which fine-tunes the projected warp path through local adjustments controlled by a radius parameter.

2.2 MediaPipe

MediaPipe is a comprehensive framework developed by Google that enables developers to create multi-modal, cross-platform applied machine learning (ML) pipelines. Designed to handle various data types – including video, audio, and time series – MediaPipe is a versatile tool for numerous applications. It is renowned for its robust collection of human body detection and tracking models, trained on some of the most extensive and diverse datasets available [15].

MediaPipe presents a hand-tracking solution to recognize and extract landmarks from RGB images. The first model detects the palm within an image and provides an accurately cropped palm image, which is then passed to the landmark model. This step reduces the need for extensive data augmentation, such as rotations, flipping, and scaling, and focuses the model's power on precise landmark localization. The Hand Landmark model processes the detected palm regions to precisely localize 21 three-dimensional hand-knuckle coordinates (x, y, z), as shown in Figure 2. This model maps coordinates even to partially visible hands and does not incorporate information regarding facial modifications [32].

The pose solution aims to detect and track the human body's skeletal structure. The model identifies 33 key landmarks on the body, including the face, shoulders, elbows, wrists, hips, knees, and ankles. This comprehensive set of landmarks allows for detailed and accurate body pose estimation, which is crucial for applications such as fitness tracking, augmented reality, and animation [4].

Figure 3 presents the tracking of MediaPipe Holistic with Hands and Pose used to detect landmarks on a frame, combining the capabilities of hand tracking, pose estimation, and face mesh into a single unified pipeline. This simultaneous tracking of the body pose, hand movements, and facial expressions provides a comprehensive understanding of human motion and interaction.

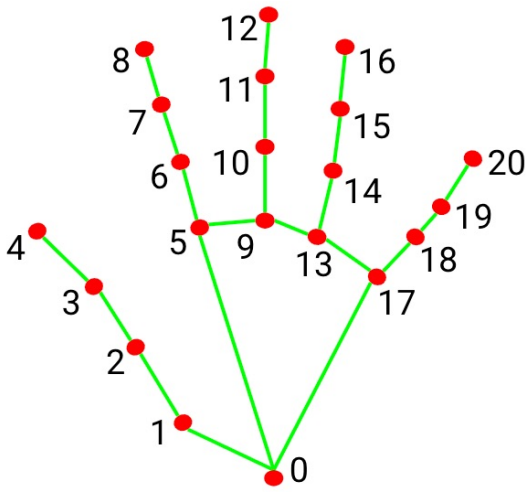

Figure 2: Hand landmarks identified by MediaPipe. Each number represents a finger joint [11].

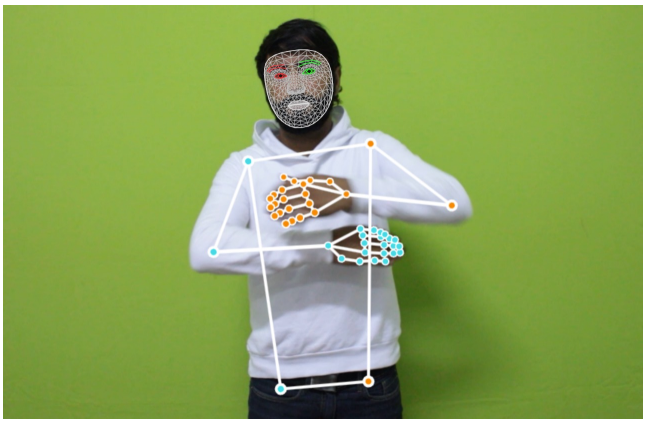

Figure 3: MediaPipe Holistic with Hands and Pose enabled to detect landmarks on a video frame.

3 RELATED WORKS

Some techniques have been developed for sign language recognition using different strategies and datasets. In early works, data gloves with embedded sensors captured intricate finger and hand movements and orientations. These gloves provide precise measurements, unaffected by external agents such as light or magnetic fields. However, data gloves are often uncomfortable and expensive, limiting their suitability for extended use, despite their accuracy [14, 17, 21].

Depth-sensing cameras like the Microsoft Kinect and Leap Motion Controller capture both RGB and depth information, allowing for more detailed gesture analysis. Adeyanju et al. [2] demonstrated the effectiveness of these devices in sign language recognition, noting their ability to capture fine-grained details of hand movements and spatial orientation. Despite that, these devices also add to the overall cost and complexity of the system.

Dynamic videos incorporating movements for signs have been widely used in computer vision-based methods. De Castro et al. [9] used RGB cameras along with the MINDS-Libras dataset and 3D CNNs to extract spatial and temporal features through 3D convolution operations, achieving an average accuracy of 91%. In addition to using RGB cameras, more advanced techniques incorporate depth sensors and infrared cameras to enhance recognition accuracy. Escobedo et al. [10] suggested a method that combines RGB-D data with texture maps to capture hand location and movement. This approach integrates multimodal data into a three-stream CNN architecture, allowing for robust feature extraction and achieving superior performance compared to traditional RGB-based methods. Their system demonstrated improvements in recognizing dynamic signs, highlighting the advantages of incorporating depth information.

Methods utilizing MediaPipe have gained attention for their real-time capabilities and ease of deployment [16, 27]. Tayade and Patil [29] conducted a study on real-time letter recognition using MediaPipe with Support Vector Machine (SVM). They utilized datasets from American, Indian, Italian, and Turkish sign languages for training and evaluation, achieving an average accuracy of 99%.

One of the common challenges in creating Sign Language Recognition (SLR) models is the lack of extensive and high-quality datasets

for certain sign languages [22]. Training models that extensively use deep learning techniques can be computationally expensive due to high data dependency. In this way, DTW is an effective strategy for this task, as it measures the similarity between temporal sequences based on their distance [3, 28]. Cheng et al. [6] utilized DTW for distance mapping in Chinese sign language recognition. They combined SVM and DTW, achieving an accuracy of 99.03% in a dataset of 11 signs, demonstrating the effectiveness of this approach for recognizing complex signs.

The literature on SLR reveals that various approaches and techniques have been employed in different contexts and scenarios. Among the related works, the use of MediaPipe stands out, demonstrating promising results in real-time gesture detection with high accuracy. These studies provide important insights and useful perspectives that contribute to the development and improvement of this work. Considering the lessons learned from related works, the proposed project aims to use MediaPipe to extract relevant gesture features and employ a DTW-based model to recognize a set of signs, thereby improving communication between deaf and hearing individuals.

4 METHODS

Sign language is not merely a collection of simple gestures; rather, it is a complex linguistic system defined by multiple parameters that allow for the encoding of a broad range of meanings. According to Brito [5], the primary parameters in sign language include hand configuration, articulation point, and movement.

Hand configuration refers to the distinct shapes that hands can assume to generate signs. The specific shapes and configurations of the hands can vary widely between different sign languages. The articulation point or location involves the space in front of the body (neutral space) or specific body regions, such as the head, waist, and shoulders, where signs are articulated. The location of the sign is crucial as it provides context and meaning. Movement is a complex parameter involving various forms and directions, including pulsing motion, movements of the finger joints, and directional movements in space. The displacement of the hands, fingers, and arms over time plays a significant role in conveying the sign's meaning.

Depending on the context, some parameters might not be necessary for interpretation. For instance, *hand orientation* refers to the direction of the palm during the sign, which can face up, down, towards the body, forward, left, or right. Orientation helps distinguish between signs that may have similar hand configurations and movements. Facial expressions and other non-manual expressions are essential for providing additional context and emphasis. They can convey emotions and differentiate between types of sentences such as affirmative, interrogative, exclamatory, and negative statements.

Figure 4 shows a flowchart of the proposed Libras recognition framework based on landmarks extraction and computing distances between multiple time series. In the first step, the video input frame is cropped using a simple region of interest (ROI) to remove potential noise caused by the borders of the videos during the recognition task. This reduction effectively trims 10% of the vertical distance on each video frame.

Next, we extract the landmarks using MediaPipe Holistic. At this stage, we execute with $min\_detection\_confidence = 0.5$ and $min\_tracking\_confidence = 0.5$, utilizing the Hands and Pose models. The hand model provides 21 three-dimensional landmarks for each hand present in a frame, while the pose model provides 33 landmarks indicating the positions of the body.

To process the results from the landmarks, we flatten all the landmarks provided by MediaPipe Holistic, resulting in a single array of coordinates. We clean up the data by filling possible null values with 0. This step is important to avoid any noise during detection, as blurred images (during rapid movements) can lead to issues in the detection [27].

With the landmarks ready, we execute the feature extraction process on the landmarks data. Figure 5 illustrates the hand landmarks identified by MediaPipe. These points are used to calculate the angles between finger joints, allowing the model to recognize different signs based on the configuration of fingers and hands. Each sign can be considered a composition of different poses, where each pose is characterized by a particular set of angles. This concept has been applied in several works in human action recognition [1, 18]. Each feature vector is composed of hands angles with several points, and pose angles. To calculate the value of an angle ω from the landmark 3D values we can use the dot product :

$$\omega = \arccos\left(\frac{\overline{BC} \cdot \overline{CD}}{|\overline{BC}||\overline{CD}|}\right) \tag{6}$$

in which $\overline{BC}$ and $\overline{CD}$ are segments composing a joint.

Once the feature vectors are defined, we use an unsupervised learning approach to provide the necessary information to the model before the tests. Leave-One-Person-Out Cross-Validation (LOPOCV) is a validation technique that is well-suited for scenarios that involve gesture recognition tasks [31]. We divide the dataset so that one individual's data is completely left out of the training set and used exclusively for testing. This process is repeated for each individual in the dataset, ensuring that each person's data is used as a test set exactly once. The model is trained on the remaining individuals' data during each iteration.

The model uses FastDTW to compute the distance between time series and perform sign recognition. This approach compares the sign to be recognized with all known signs and measures the distance among them. The FastDTW algorithm is executed with the radius parameter set to 1, which defines the size of the neighborhood when expanding the path. After computing all distances, the closest known sign name is used in the output of the task.

5 RESULTS

We evaluated the proposed method on MINDS-Libras [22] and INCLUDE-50 [26] datasets. The MINDS-Libras dataset was used to optimize the model, including adjustments to the region of interest in the videos and configuring the model to better recognize the signs used. Once the model was established, it was applied to both datasets to test in different scenarios and compare to the state-of-the-art results.

The MINDS-Libras dataset [22] consists of 1,200 data sequences distributed into 20 classes. Each class represents a distinct sign from Libras, including both static and dynamic signs. The signs were captured using two types of sensors: a Canon EOS Rebel t5i DSLR camera and a Microsoft Kinect v2 sensor. As a result, each

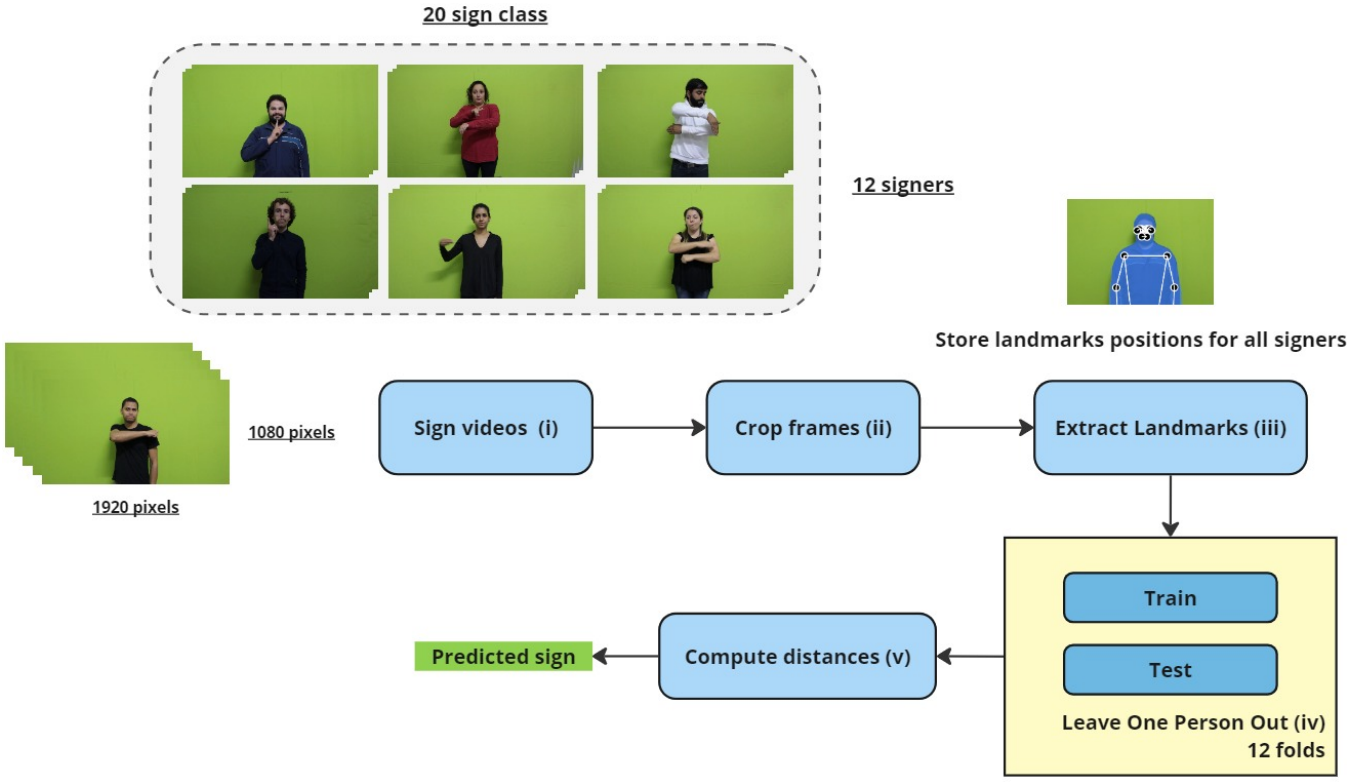

Figure 4: Outline of the proposed method for Libras recognition.

recording contains a 1920 × 1080 RGB video (captured by the DSLR camera) and 640 × 480 RGB-D data (captured by the Kinect sensor), both recorded at 30 frames per second (fps). However, we used only the full HD RGB data in this work.

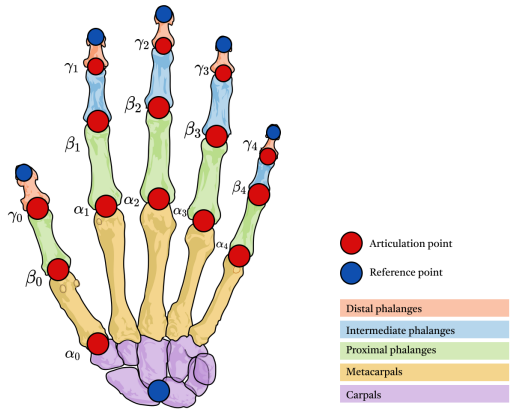

Figure 5: The feature extracted from the hand: joint angles and fingertip positions. The blue points indicate the fingertip positions on which the 3D displacements are computed. The red points indicate the joints on which the angles are computed.

We used overall accuracy and F1-score as performance metrics to evaluate our model. Accuracy is the proportion of true positive (TP) and true negative (TN) results among the total number of cases examined. It is calculated using the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

in which FP and FN represent false positives and false negatives, respectively. The F1-score is the harmonic mean of precision and recall, balancing the two metrics. It is calculated as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positives, while Recall is the ratio of correctly predicted positive observations to all observations of the actual class.

Table 1 shows the models' performance metrics. The results in Table 1 were obtained using the Leave-One-Person-Out Cross-Val-i-da-tion (LOOCV) technique, a method proposed by De Castro et al. [9]. The signers from the dataset were divided into 12 groups, with 11 for validation and 1 for testing. This approach is applied to achieve an accuracy of 0.86 ± 0.08, using only RGB data as input to the network and without employing deep learning or data augmentation to improve the results. Additionally, the results from De Castro et al. [9] and Passos et al. [20], both of which use data augmentation, are presented for comparison. Despite not using data augmentation, our approach continues to yield competitive

Table 1: Comparison of methods for recognizing signs using MINDS-Libras dataset.

Method	Aug	Dataset	Accuracy	F1-Score
De Castro et al. [9]	Yes	MINDS-Libras	0.91 ± 0.07	0.90
Passos et al. [20]	Yes	MINDS-Libras	0.85 ± 0.02	-
Ours	No	MINDS-Libras	0.86 ± 0.08	0.87

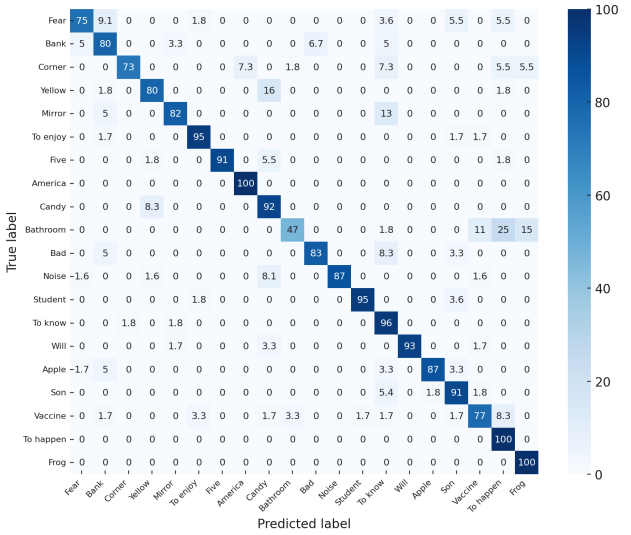

Figure 6: Confusion matrix across all runs on the MINDS-Libras dataset.

results, primarily because it does not require a large amount of newly produced information for training.

The confusion matrix in Figure 6 illustrates the performance of our model across all runs on the MINDS-Libras dataset. The matrix shows that some signs have a higher rate of confusion. For example, "Yellow" was misclassified as "Candy" in 0.16 of the cases, and "Bathroom," which has the lowest accuracy at 0.47, was misclassified as "To Happen" in 0.25 of the cases. These areas highlight where additional contextual features, such as facial expressions and the position of hands relative to the face, could improve recognition accuracy.

Table 2 shows the best case for precision, recall, and F1-score achieved by Signer 12 using the LOOCV method. The model demonstrates an overall accuracy of 0.96 achieving perfect scores in precision, recall, and F1-score for 16 classes, indicating that it recognized these signs without any false positives or false negatives.

However, certain signs, such as "Bank", "Frog", "Bathroom", and "Bad", show lower precision and recall values. These lower precision and recall values can be attributed to the similarity in hand configurations or movements with other signs. For example, "Bank" is confused with "Bad" due to the high similarity between the signs and the fact that the model considers the angles formed by the trunk and hands but not the position of the hand relative to the face, which hinders the precise identification of certain signs. Another example is "Frog" being confused with "Bathroom", which could have been

Table 2: Best case for precision, recall, and F1-score for each class (Signer 12) using the MINDS-Libras dataset with the LOOCV method.

	Precision	Recall	F1-Score	Support
To happen	1.00	1.00	1.00	5
Bank	0.71	1.00	0.83	5
Mirror	1.00	1.00	1.00	5
To know	1.00	1.00	1.00	5
Five	1.00	1.00	1.00	5
Apple	1.00	1.00	1.00	5
Corner	1.00	1.00	1.00	5
Bathroom	1.00	0.60	0.75	5
Student	1.00	1.00	1.00	5
Frog	0.71	1.00	0.83	5
To enjoy	1.00	1.00	1.00	5
To know	1.00	1.00	1.00	5
Fear	1.00	1.00	1.00	5
America	1.00	1.00	1.00	5
Will	1.00	1.00	1.00	5
Yellow	1.00	1.00	1.00	5
Vaccine	1.00	1.00	1.00	5
Son	1.00	1.00	1.00	5
Noise	1.00	1.00	1.00	5
Bad	1.00	0.60	0.75	5
Accuracy			0.96	100

avoided if the model had considered facial expressions during sign identification.

To compare our model with the proposed model for validating the MINDS dataset, it was necessary to use the same separation technique proposed in MINDS, with a random split of 75% of the dataset used for training and 25% used for validation. An important point to highlight about this technique is that it can cause overfitting in the model, which can hinder the development of the model and its ability to generalize. This issue is mentioned in De Castro et al. [9], which discusses overfitting and proposes the previously mentioned LOOCV technique. Table 3 shows the results using the random split; our model achieved an accuracy of 0.98 ± 0.01, which is superior to that obtained by the model proposed in Rezende et al. [22]. It is important to note that our model does not use data augmentation techniques, which are often used to artificially increase the size and variability of the training data to improve model performance.

The INCLUDE-50 dataset, compiled by Sridhar et al. [26], comprises 263 distinct classes of signs in Indian Sign Language (ISL). These classes are organized into 15 categories including clothes, colors, adjectives, and pronouns. Each signer performs each sign

Table 3: Comparison of methods for recognizing Libras signs using a 75% training and 25% testing split utilizing MINDS dataset.

Method	Aug	Dataset	Accuracy	F1-Score
Rezende et al. [22]	Yes	MINDS-Libras	0.93 ± 0.02	0.93
Ours	No	MINDS-Libras	0.98 ± 0.01	0.98

Table 4: Comparison of methods for recognizing signs using INCLUDE-50 dataset.

Method	Aug	Dataset	Accuracy
De Castro et al. [9]	Yes	INCLUDE-50	0.95
Sridhar et al. [26]	Yes	INCLUDE-50	0.94
Sridhar et al. [26]	No	INCLUDE-50	0.74
Ours	No	INCLUDE-50	0.90

2 to 6 times, resulting in 4,287 videos. For quick evaluation, the INCLUDE-50 subset, also created by Sridhar et al. [26], includes 50 sign categories with 958 videos. In our study, we utilize this subset to validate our methodology. Sridhar et al. [26] previously defined the videos that make up the training and test sets. We employed the same division proposed by them.

The results for the INCLUDE-50 dataset are shown in Table 4. The proposed model achieved an accuracy of 0.90, demonstrating the effectiveness of our approach in different sign languages and validating its applicability in various scenarios. The comparison with other methods, such as the work by De Castro et al. [9] and Sridhar et al. [26], shows that our approach performs competitively, particularly in terms of accuracy, without relying on deep learning techniques, especially when compared to the method proposed by Sridhar et al. [26] without data augmentation, achieving an improvement of approximately 21.5%.

These findings highlight the robustness and generalizability of our proposed method across different datasets and sign languages. The use of data augmentation has a significant impact on the performance of models. For instance, Sridhar et al. [26] achieved an accuracy of 0.94 with augmentation, but only 0.74 without it. Despite not using data augmentation, our model achieved a competitive accuracy of 0.90, showcasing the effectiveness of our approach. Further studies can build upon this work to explore more sophisticated models and techniques, aiming for even higher accuracy and broader applicability in real-world scenarios.

6 CONCLUSION

This research presents an unsupervised method for sign recognition. The model that uses FastDTW, together with angle measurement, demonstrated good generalization capacity across two datasets, without the use of data augmentation. Our approach uses visual information to describe signs without the analysis of depth sensors or gloves, facilitating its dissemination to different languages and datasets.

For future work, we plan to incorporate data augmentation techniques, which have shown significant improvements in accuracy. One possible model enhancement is to include the position of the hand relative to the face as a feature to be measured, as well as identifying contextual cues such as facial expressions and other parts of the body. This can be essential for understanding the sign and further improving the model's accuracy. Another aspect we intend to address is the weighting of angles, allowing more important angles to have greater influence, thereby aiding the model's identification process.

ACKNOWLEDGMENTS

The authors thank the Pontifícia Universidade Católica de Minas Gerais – PUC-Minas, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (Grant PROAP 88887.842889/2023-00 – PUC/MG, Grant PDPG 88887.708960/2022-00 – PUC/MG - Informática, Grant STIC-AMSUD 88887.878869/2023-00 and Finance Code 001), the Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Grants 407242/2021-0, 306573/2022-9 and 442950/2023-3), and Fundação de Apoio à Pesquisa do Estado de Minas Gerais – FAPEMIG (Grant APQ-01079-23).

REFERENCES

[1] Sunusi Bala Abdullahi and Kosin Chamnongthai. 2022. American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach. *IEEE Access* 10 (2022), 15911–15923.
[2] Ibrahim Adepoju Adeyanju, Oluwaseyi Olawale Bello, and Mutiu Adesina Adegboye. 2021. Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications* 12 (2021), 200056.
[3] Nikolaos Arvanitis, Evangelos Sartinas, and Dimitrios Kosmopoulos. 2023. Procrustes-DTW: Dynamic Time Warping Variant for the Recognition of Sign Language Utterances. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 1–5.
[4] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204* (2020).
[5] Lucinda Ferreira Brito. 2010. *Por uma gramática de línguas de sinais.* TB-Edições Tempo Brasileiro.
[6] Juan Cheng, Fulin Wei, Yu Liu, Chang Li, Qiang Chen, and Xun Chen. 2020. Chinese Sign Language Recognition Based on DTW-Distance-Mapping Features. *Mathematical Problems in Engineering* 2020, 1 (2020), 8953670.
[7] Bruno Costa, Jean Freire, Hamilton Cavalcante, Márcia Homci, Adriana Castro, Raimundo Viégas Jr, Bianchi Meiguins, and Jefferson Morais. 2017. Fault Classification on Transmission Lines Using KNN-DTW. 174–187. https://doi.org/10.1007/978-3-319-62392-4_13
[8] Diego RB da Silva, Tiago Maritan U Araujo, Thais Gaudencio do Rêgo, and Manuella Aschoff Cavalcanti Brandão. 2020. A Two-Stream Model Based on 3D Convolutional Neural Networks for the Recognition of Brazilian Sign Language in the Health Context. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 5–12.
[9] Giulia Zanon De Castro, Rubia Reis Guerra, and Frederico Gadelha Guimarães. 2023. Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps. *Expert Systems with Applications* 215 (2023), 119394.
[10] Edwin Escobedo, Lourdes Ramirez, and Guillermo Camara. 2019. Dynamic Sign Language Recognition Based on Convolutional Neural Networks and Texture Maps. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 265–272. https://doi.org/10.1109/SIBGRAPI.2019.00043
[11] Google. 2024. MediaPipe Pose. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker
[12] Rohit J Kate. 2016. Using dynamic time warping distances as features for improved time series classification. *Data mining and knowledge discovery* 30 (2016), 283–312.
[13] Deep R. Kothadiya, Chintan M. Bhatt, T. Saba, A. Rehman, and Saeed Ali Omer Bahaj. 2023. SIGNFORMER: DeepVision Transformer for Sign Language Recognition. *IEEE Access* 11 (2023), 4730–4739. https://doi.org/10.1109/ACCESS.2022.3231130
[14] Boon Giin Lee and Su Min Lee. 2017. Smart wearable hand device for sign language interpretation system with sensors fusion. *IEEE Sensors Journal* 18, 3 (2017), 1224–1232.
[15] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
[16] Marc Marais, Dane Brown, James Connan, and Alden Boby. 2022. Improving signer-independence using pose estimation and transfer learning for sign language recognition. In *International Advanced Computing Conference*. Springer, 415–428.
[17] Syed Atif Mehdi and Yasir Niaz Khan. 2002. Sign language recognition using sensor gloves. In *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, Vol. 5. IEEE, 2204–2206.
[18] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2014. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 24–38.

[19] Luiza Maria Borges Oliveira. 2012. *Cartilha do Censo 2010 – Pessoas com Deficiência.* Secretaria de Direitos Humanos da Presidência da República, Brasília.
[20] Wesley L Passos, Gabriel M Araujo, Jonathan N Gois, and Amaro A de Lima. 2021. A gait energy image-based system for Brazilian sign language recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers* 68, 11 (2021), 4761–4771.
[21] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications* 164 (2021), 113794.
[22] Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. 2021. Development and validation of a Brazilian sign language database for human gesture recognition. *Neural Computing and Applications* 33, 16 (01 Aug 2021), 10449–10467. https://doi.org/10.1007/s00521-021-05802-4
[23] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.
[24] Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* 11, 5 (2007), 561–580.
[25] Maria Fernanda Neves Silveira de Souza, Amanda Miranda Brito Araújo, Luiza Fernandes Fonseca Sandes, Daniel Antunes Freitas, Wellington Danilo Soares, Raquel Schwenck de Mello Vianna, and Árlen Almeida Duarte de Sousa. 2017. Principais dificuldades e obstáculos enfrentados pela comunidade surda no acesso à saúde: uma revisão integrativa de literatura. *Revista Cefac* 19 (2017), 395–405.
[26] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. 2020. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*. 1366–1375.
[27] Barathi Subramanian, Bekhzod Olimov, Shraddha M Naik, Sangchul Kim, Kil-Houm Park, and Jeonghong Kim. 2022. An integrated mediapipe-optimized GRU model for Indian sign language recognition. *Scientific Reports* 12, 1 (2022), 11964.
[28] Jimin Tan, Jianan Yang, Sai Wu, Gang Chen, and Jake Zhao. 2021. A critical look at the current train/test split in machine learning. *ArXiv preprint ArXiv:2106.04525* (2021).
[29] Akshit Tayade and Swapnil Patil. 2021. Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning. *International Journal of Research Publication and Reviews* 2, 5 (2021), 9–17. https://doi.org/10.13140/RG.2.2.32364.03203
[30] Ankita Wadhawan and Parteek Kumar. 2020. Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications* 32 (2020), 7957 – 7968. https://doi.org/10.1007/s00521-019-04691-y
[31] Tzu-Tsung Wong. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* 48, 9 (2015), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009
[32] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214* (2020).