A Cross-Dataset Study on the Brazilian Sign Language Translation

Amanda Hellen de Avellar Sarmento, Moacir Antonelli Ponti
ICMC - Universidade de São Paulo
São Carlos, SP, Brazil
amanda.avellar@usp.br, ponti@usp.br

Abstract

Signed communication is an important form of natural language, often less studied, but still relevant. The main question we address in this paper is how to translate Brazilian Sign Language (LIBRAS) implementing Deep Learning networks with limited data availability. Previous studies often use a single dataset, in most cases collected by the authors themselves. We claim a cross-dataset approach would be more adequate to evaluate real-world scenarios. We investigate two methods based on spatial feature extraction. The first one uses pre-trained Convolutional Neural Networks (CNN) and the second one Body Landmark Estimation (skeleton information). A Long Short-Term Memory (LSTM) network is responsible for the sign classification. Our contribution encompasses data curation, alongside providing general guidelines for enhanced generalization.

1. Introduction

Spoken and signed communication are both considered natural languages. In particular, sign languages use a visual-manual approach to convey meaning, having their own grammar and lexicon, and forming the core of local Deaf communities [21]. There are more than 200 sign languages registered, going back from the 5th century BC [4]. Yet, historically such languages were marginalized, with limited documentation consisting of manual alphabets (fingerspelling systems) to transfer words from a spoken language to a sign language, and not of the language itself [14]. Currently in Brazil more than 10 million people have some degree of disabling hearing loss. When it comes to education, nearly half of the such population does not reach high school [8, 3]. Computational approaches that allow processing, translation, and interaction using sign language are essential to better include the deaf population.

In this paper we focus on recognizing and translating signs from the Brazilian Sign Language (LIBRAS). The main difficulty is the lack of data, both in terms of data availability and variability. Typically, researchers build their own video datasets in order to study the problem [17, 22, 5, 2, 9, 24]. The limitations of such approach include having models that deal with a small vocabulary (number of words/categories) as well as a dependency on having a controlled environment [16].

The lack of a common training dataset across studies leads to results that hardly compare against one another [1, 7], raising questions about the models' generalization to real-world conditions.

This paper's contribution lies on gathering diverse and reliable sources, covering multiple regions from Brazil, which to the best of our knowledge, has not yet been done. Consequently, we enable a more comprehensive study and foster future research by making the data publicly available, while providing comparable results.

Additionally, we take into consideration the different formats of sign language data, given there are different ways of collecting it. The options range from sophisticated devices such as sensory gloves, depth cameras and optical hand tracking [20, 25], to simpler devices such as standard cameras, *e.g.*, smartphones' built-in cameras. Bearing in mind the model requires consistent input data format throughout all steps, standard cameras offer a practical advantage in a real-world use. Despite not being tailored for sign language, they are accessible and cost-effective for potential users of the model.

We investigate methods for frame sampling, network models and different spatial feature extraction. In particular we employed either CNNs or Body Landmark Estimation (skeleton information) [6, 13], both followed by a LSTM classification network.

2. Data Sources

In order to analyze the models' generalizability in a more realistic scenario, we searched for multiple, reliable data sources. We prioritised gathering data recorded by standard camera devices. Ultimately, we collected data from four distinct sources. We used the largest dataset (UFPE) as reference to define the subsequent signs to be collected.

Federal University of Pernambuco (UFPE): V-LIBRASIL dataset, referred to as UFPE source, was developed as part of Rodrigues' master's thesis [18]. The lack of a robust LIBRAS dataset and further Deep Learning (DL) studies was acknowledged and in order to bridge this gap, the researcher made it available for download[1];
Federal University of Viçosa (UFV): LIBRAS-Portuguese Dictionary dataset, referred to as UFV source, developed by Projeto "Inovar Mais" with a pedagogical focus for students and teachers from the university. The dataset is public but requires direct contact to access[2];
National Institute of Deaf Education (INES): LIBRAS Dictionary dataset, referred to as INES source, developed by the National Institute of Deaf Education. The Institute has an extensive content production, such as videos in LIBRAS, distributed to educational systems. It is publicly available[3], but not direct download, requiring scraping;
SignBank: The SignBank dataset, referred to as SignBank source, was developed by the Federal University of Santa Catarina. Along with the videos, linguistic aspects such as semantic field, word syntax, dominant hand configuration, *etc*., are also present on the website[4]. Its goal is to make the data available to national and international deaf communities, as well as to serve as a linguistic research source. The data is also not directly available for downloading, so we followed the same process as for the INES data source.

The integration of the data sources involved many challenges. After the collection phase, the labels had to be extracted, *e.g*., for the UFPE source the label was embedded on the video, requiring the use of Optical Character Recognition (OCR) techniques. We then cleaned the labels, replacing separator symbols for white spaces, removing line breaks, numbers, punctuation, and other symbols. All letters were converted to lowercase for consistency. Subsequently, the labels were reasonably standardized. Besides that, we cleaned the data, discarding videos that were either empty or containing rendering errors. The INES source contained a higher number of videos with this type of error.

Table 1 shows the number of categories (signs/words), number of observations per category, and the total number of observations per source. The Cross-Dataset contains videos of isolated signs and with by-request access[5].

The number of observations per category is scarce. Therefore, to address this limitation, only categories present in all four datasets were selected, ensuring that each category has at least four observations. It is worth to mention that each data source may have labelled the same signs differently, due to synonyms in the Portuguese language. Additionally, given regional linguistic differences, the datasets are composed of different words as well, leading to a small intersection. As a result, the final dataset has 49 categories, with 6 observations per category on average, and a total of 313 observations.

Source	# of distinct categories	# of obs per category	# of total obs
UFPE	1396	1 to 7	4221
UFV	1004	1 to 2	1029
INES	237	1 to 2	282
SignBank	485	1	485
Total	2098	1 to 12	6017

Table 1: Number of distinct categories, observations per category and total observations per source.

The dataset categories are (translated to English): pineapple, to accompany, to happen, to wake up, to add, tall, friend, year, before, erase, to learn, air, beard, boat, bicycle, goat, ox, ball, bag, hair, to fall, box, calculator, wedding, horse, onion, beer, to arrive, flip-flops, coconut, rabbit, to eat, to compare, to buy, computer, to destroy, day, to decrease, elephant, elevator, school, to choose, to forget, flute, flower, watermelon, to mix, to swim, roller skates.

3. Method

Our approach includes preprocessing the videos, sampling frames, splitting the sets, performing data augmentation and resizing it. Then we proceed to extract spatial features and use them as input to the LSTM network, which performs the classification. At last, the performance is evaluated. For the extraction step we explored two methods. The first one explores three pre-trained CNN models and the second one Pose and Hand Landmarks. The latter one involves estimating key body locations on videos to analyze posture and movements.

3.1. Preprocessing

Frame selection: we randomly selected 15 frames from each video using a normal distribution, which showed to be better than the uniform on the experiments. We believe this is because the videos were already preprocessed and temporally centered.
Train, val and test sets: The training, validation and test sets split is 70%, 15% and 15% respectively. The validation and test observations were exclusively from UFPE dataset, as it is the only source to have at least three observations per category. Thus, all sources were present in the training set.
Data augmentation: each video was augmented up to 20 times using a combination of random transformations: horizontal flipping, rotation, translation, centered crop, brightness and contrast adjustment. In addition, for each new

[1] https://libras.cin.ufpe.br/
[2] https://sistemas.cead.ufv.br/capes/dicionario/
[3] https://www.ines.gov.br/dicionario-de-libras/
[4] https://signbank.libras.ufsc.br/pt
[5] https://github.com/avellar-amanda/LIBRAS-Translation/

instance, a different sample of frames was drawn, adding more variance to the augmented data. We performed experiments using augmented data up to 1, 5, 10, 15 and 20 times, and the latter option demonstrated better results.
Resizing: frames were resized to 640 in width and 480 in height. Such resizing previous to the landmark estimation is crucial, given that it affects the landmarks' 3D coordinates.

3.2. CNN Feature Extraction

Three pre-trained CNN models: i) MobileNetV2 [21], ii) InceptionResNetV2 [23] and iii) ResNet50V2 [11], since those were shown to be good general-purpose feature extractors [12, 15]. Each video frame was resized to 224×224 for input. A global average pooling was applied on the last convolutional block to result in 1280, 1536 and 2048 features respectively.

3.3. Pose and Hand Landmarks

The MediaPipe library [10] was used to obtain the person's landmarks [19]. MediaPipe was designed to facilitate the development of applications involving real-time media analysis by providing modular building blocks and an efficient data processing pipeline. Additionally, it is open source, facilitating customization.

The Holistic model was used, as it estimates landmarks of the pose, hands and face at the same time. From the outputs, we experimented with landmarks from the pose (33), face (468) and hands (21 for each hand), but the face landmarks led to worse performance, thus only pose and hands landmarks were further explored. Figure 1 shows an example of a sampled frame from the sources with the respective landmarks below, drawn on a white background for illustration purposes. Note that the landmarks help focusing on important features, avoiding spurious features such as background, color, lighting and identity of the signer.

3.4. Landmark-based Feature Extraction

From 3D landmarks, we computed 52 angles between adjacent landmark connections of each hand, as well as 38 distances between specific pose landmarks (aiming the LIBRAS translation). Both are combined in a final feature vector with 90 features. Figure 2 illustrate the features. In Figure 2a, the green curved line illustrates distances between landmark pairs. In Figure 2a, green lines represent connections between the landmarks and black curved lines illustrate the angle between some adjacent connections.

Prior to the distance computation, the pose landmarks were standardized by subtracting the center pose and dividing by the maximum pose size. The center was considered as the point between the shoulders' landmark. The maximum pose size was defined as the longest distance between each landmark and the pose center. With this we standardized the translation and scale of the landmarks.

3.5. LSTM Networks for Classification

The LSTM networks had the following architecture: i) input layer with 15 frames and feature dimension according to the feature extractor, ii) normalization layer, iii) mask layer with value zero, iv) one LSTM layer with 128, 256 or 512 units, ReLu as the activation function and L1 regularizer with 0.001 factor, v) dropout layer with 0.4 factor and vi) classification layer with Softmax as the activation function. We employed: i) sparse categorical cross entropy as the loss function, ii) AdamW, with learning rate of 0.0001 and weight decay of 0.005, as the optimizer. All models were evaluated with accuracy and top-5 accuracy, which are adequate since the dataset is not imbalanced. To define the fixed parameters a Random Search was performed with the MobileNetV2-LSTM model.

4. Experimental Results

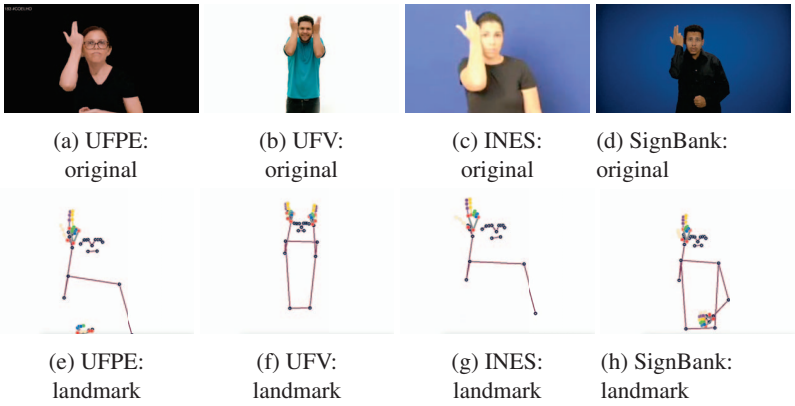
The networks were trained for up to 200 epochs. An early stopping criteria was set with a patience of 20 and monitored with the validation accuracy. Once the training was complete, the best weights were restored.

Table 2 presents all the results divided into three parts (for which only the best three outcomes according to test accuracy and test top-5 accuracy are displayed): i) CNN-LSTM configurations, ii) Landmark-based feature extractor with LSTM configurations, and iii) the former experiment conducted with a subset of categories.

The best CNN model achieved 10% test accuracy and 95% train accuracy. We noticed the model overfits, failing to capture meaningful patterns. Among the possible irrelevant patterns the model might have learned, the most noticeable to human eyes may be the colors of background or clothes present in the original videos, as illustrated in Figure 1. The background and clothing color were, respectively: i) black and black for the UFPE source, ii) white and blue for the UFV source, iii) lilac and gray for the INES source and iv) blue and black for the SignBank source. In summary, although there was some learning (the random accuracy would be approximately 2%), none of the models were able to properly generalize to the test set.

The increase in performance is significant when replacing the CNN backbones with a Landmark-based feature extractor, reaching 41% test accuracy in the best case. Analyzing the top-5 accuracy, we notice that among the predicted categories with the highest probability, the correct one was in the top-5 approximately 75% of the time, which comes closer to a reasonable result.

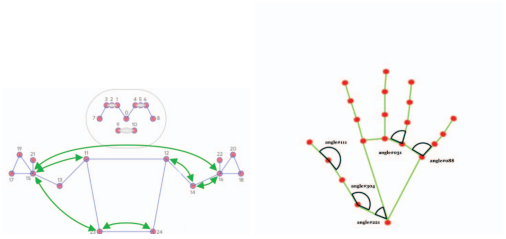
Based on the previous experiment, the miss-classified categories were analysed. We observed that certain categories exhibited a high internal variance, meaning that the signs performed within these categories were sometimes significantly different from one another.



(a) UFPE: original (b) UFV: original (c) INES: original (d) SignBank: original

(e) UFPE: landmark (f) UFV: landmark (g) INES: landmark (h) SignBank: landmark

Figure 1: Example of a sampled frame (sign "rabbit") from different sources vs. drawn landmarks on a white background.



(a) Distance between pose landmarks. (b) Angle between hand connections.

Figure 2: Example of the extracted features.

Feature extractor	# units LSTM	# categ.	Acc.	Top 5 acc.
InceptionResNetV2	128	49	10.2	26.5
MobileNetV2	256	49	08.2	22.4
InceptionResNetV2	512	49	08.2	22.4
Landmark-based	256	49	40.8	75.5
Landmark-based	512	49	40.8	67.3
Landmark-based	128	49	38.8	65.3
Landmark-based	512	33	66.7	93.9
Landmark-based	256	33	63.6	87.9
Landmark-based	128	33	60.6	90.9

Table 2: Comparative evaluation of the proposed methods.

The same way spoken languages have dialects, it is natural for sign language to have regional variations as well. However, given the limited number of observations per category, a few deviations can negatively impact the model's performance. Therefore, we selected categories where at least 50% of the observations didn't exhibit high variance. For the final experiment we considered a subset of 33 categories. The removed categories were: air, to fall, calculator, to arrive, to decrease, elephant, elevator, to forget, watermelon, to happen, beard, wedding, beer, day, swim, roller skates. As a result, we obtained approximately 66% test accuracy and 94% test top-5 accuracy.

Additionally, we performed an external validation, in order to check whether the model was able to generalize to a data source not seen during the training process. Even when the UFPE source (with the largest number of observations) was not present in the training set, the model still obtained around 54% test accuracy.

5. Conclusion

Our results indicate that Brazilian Sign language translation remains an open problem. For training and validation on the same dataset, previous studies observed high accuracy (around 80%) using, for example, a combination of convolutional and recurrent layers. However, when a cross-dataset is used, the same approach does not generalize.

We show that Landmark-based features may be a better option towards this problem, and that sampling the frames using a normal distribution, *i.e.*, central frames are more likely to be selected, may improve the learning. Moreover, sampling the frames based on a random process introduces greater diversity to the augmented data, enabling the capture of different aspects that may have been missed in previous samples. Finally, despite having a limited number of observations per category as the starting point, this study was able to provide valuable insights and results that are closer to a more realistic implementation of LIBRAS translation.

Future work includes scaling up the data collection from various sources and implementing a video similarity-based model to identify outliers in each category, in addition to exploring different DL methods. Our ultimate goal is to translate LIBRAS, encompassing entire sentences and capturing the contextual meaning, rather than solely focusing on word-by-word translations.

References

[1] IA Adeyanju, OO Bello, and MA Adegboye. Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, 12:200056, 2021.

[2] Sílvia Grasiella Moreira Almeida, Frederico Gadelha Guimarães, and Jaime Arturo Ramírez. Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors. *Expert Systems with Applications*, 41(16):7259–7271, 2014.

[3] Assembleia Legislativa do Estado de São Paulo. International sign language day seeks to promote inclusion of deaf people (in portuguese), Setembro 2021.

[4] H-Dirksen L Bauman. *Open your eyes: Deaf studies talking*. U of Minnesota Press, 2008.

[5] Lourdes Ramirez Cerna, Edwin Escobedo Cardenas, Dayse Garcia Miranda, David Menotti, and Guillermo Camara-Chavez. A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications*, 167:114179, 2021.

[6] Jen-Li Chung, Lee-Yeng Ong, and Meng-Chew Leow. Comparative analysis of skeleton-based human pose estimation. *Future Internet*, 14(12):380, 2022.

[7] João Marcos Cardoso da Silva, Pedro Martelleto Bressane Rezende, and Moacir Antonelli Ponti. Detecting and mitigating issues in image-based covid-19 diagnosis. In *Workshop on Healthcare AI and COVID-19*, pages 127–135. PMLR, 2022.

[8] Lívia Dias, Ruth Mariani, Cristina MC Delou, Erika Winagraski, Helder S Carvalho, and Helena C Castro. Deafness and the educational rights: A brief review through a brazilian perspective. *Creative Education*, 2014, 2014.

[9] E Escobedo-Cardenas and G Camara-Chavez. A robust gesture recognition using hand local data and skeleton trajectory. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1240–1244. IEEE, 2015.

[10] Google for Developers. Compose on-device ML in minutes, 2023.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[13] MediaPipe GitHub Holistic. Mediapipe holistic, Janeiro 2023.

[14] Carol Padden. Sign language geography. *Deaf around the world: The impact of language*, pages 19–37, 2010.

[15] Moacir A Ponti, Fernando P dos Santos, Leo SF Ribeiro, and Gabriel B Cavallari. Training deep networks from zero to hero: avoiding pitfalls and going beyond. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 9–16. IEEE, 2021.

[16] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[17] Tamires Martins Rezende. Automatic recognition of libras signs: development of the minds-libras database and convolutional network models (in portuguese). 2021.

[18] Ailton José Rodrigues. V-LIBRASIL: a dataset with signs of the brazilian sign language LIBRAS (in portuguese). Master's thesis, Universidade Federal de Pernambuco, 2021.

[19] Karl Rohr. *Landmark-based image analysis: using geometric and intensity models*, volume 21. Springer Science & Business Media, 2001.

[20] Zinah Raad Saeed, Zurinahni Binti Zainol, BB Zaidan, and AH Alamoodi. A systematic review on systems-based sensory gloves for sign language pattern recognition: An update from 2017 to 2022. *IEEE Access*, 2022.

[21] Wendy Sandler and Diane Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006.

[22] Diego Ramon Bezerra Da Silva. A multistream architecture based on deep learning for LIBRAS sign recognition for the health context (in portuguese). Master's thesis, Universidade Federal da Paraíba, João Pessoa, 2020.

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[24] Johann Felipe Voigt. Deep learning for the recognition of hand guestures using images and skeletons applied to LIBRAS (in portuguese). Master's thesis, Universidade Federal de Alagoas, Maceió, 2018.

[25] Ankita Wadhawan and Parteek Kumar. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28:785–813, 2021.