A Brazilian Sign Language Recognition System

Jéssica Ramos
jessicalfr@ufmg.br

Abstract

More than 70 million people worldwide are deaf, and most use sign languages to communicate. Since sign languages usually have a completely different structure from the spoken languages in a country or region, there is a language barrier for the deaf and speech-impaired community in daily interactions. Modern computer vision systems can perform tasks of sign language recognition and serve as a translation system for communication. This paper presents a simplified version of a sign language recognition system for the Brazilian Sign Language (Libras) alphabet. The final program captures images from a webcam and translates the signs for each letter into written letters in real-time. The CNN classification model achieved an accuracy of 99.81%. Video presentation: https://youtu.be/3LrPsbP9OD8

1. Introduction

There are more than 70 million deaf people worldwide using more than 300 different sign languages according to the World Federation of the Deaf [1]. In Brazil, more than 2.3 million who don't listen completely or have great difficulty hearing [2]. Since sign language has different structures from spoken languages, there is a language barrier for the deaf and speech-impaired community in daily interactions. Sign language recognition is a possible solution for this problem.

Sign language involves the usage of different parts of the body: fingers, hands, arms, head, body, and facial expressions. From these body parts, we can extract five main parameters from sign language: hand shape, palm-orientation, movement, location, and non-manual signals like facial expression and mouthing. These parameters determine the meaning of each word and phrase during communication and can help systems in the classification of words and the translation of entire phrases.

A sign language recognition system can be vision-based or sensor-based. Sensor-based systems use different types of sensors to infer movement, hand and body positions, and translate them to text. Gloves with sensors inside of them can be used to acquire data of the palms and fingers' movements, for example. Vision-based systems use image, video, infra-red sensors, skeleton information, and flow information to recognize the gestures. These approaches are more easily adapted to real-life settings as access to cameras is easier.

There are various challenges in automatic sign language recognition, such as inter and intra-subject variability, illumination conditions, partial occlusions, different points of view and resolutions, and background artifacts. Also, most of the efforts in building sign language recognition systems focus on recognizing only one sign at a time [3], even though to support communication a system would need to make continuous translation, i.e. multiple signs in a sequence.

With the progress with deep learning in recent years, much work has been developed in sign language recognition [3]. Deep learning allows computers to learn and represent data with multiple levels of abstraction and implicitly capture complex structures of large-scale data. Most vision-based models use Convolutional Neural Networks (CNNs) for feature extraction in images [4] [5]. In the case of video inputs, Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), and Gated Recurrent Unit Networks (GRUs) capture temporal information. Some models combine multiple approaches to improve model accuracy [6] . There is also work combining hand-crafted methods with deep learning methods or traditional classifiers [7].

This work presents a simplified version of a sign language recognition system that uses static images to classify letters of the Brazilian Sign Language (Libras) alphabet. Section 2 discusses some recent work with sign language recognition using deep learning techniques and work directed to Libras recognition. Section 3 describes the methodology. Section 4 describes and discusses the results of the experiments. Section 5 concludes the findings and discusses possible approaches for future work.

2. Related Work

There are numerous works in sign language recognition for a variety of different sign languages, using both video and static images in the case of vision-based systems. Wad-

hawan and Kumar [5] use a deep learning-based CNN to classify 100 unique signs from static images of Indian Sign Language (ISL). The classes considered the alphabet, 0-9 digits, and 67 commonly used words, and the dataset was composed of 35,000 images in total. The proposed approach was evaluated in their dataset — using test data — and achieved an accuracy of 99.72% on colored images and 99.90% on grayscale images. Rastgoo et al. [7] proposed a cascaded model composed of two steps: hand detection and sign recognition. They use a proposed dataset of 10,000 RGB videos from 100 Persian sign words. The hand detection used a Single Shot Detector (SSD) model using five online sign dictionaries videos. Then the detected hands are fed to a CNN (ResNet50 [8]) to extract high-level features. These features are fused with Extra Spatial Hand Relation (ESHR) features and hand pose features and fed to an LSTM model. Finally, a fully connected layer is used for word sign classification. The model achieved an accuracy of 86.32% on isoGD dataset [9].

In the specific case of Libras, Furtado and Oliveira [10] developed a system for automatic recognition of the letters in the alphabet with videos as inputs. They built their dataset and extracted 87,000 frames from the videos representing the 26 letters and two controls signals — one to represent space between characters and the other to represent deleting the previous signal. The frames were processed by removing the background and using a filter to identify borders. The classification was performed using the grayscale borders image as input to a CNN architecture. The Inception Network [11] was used as a base model and refined to the specific task. The test dataset achieved an accuracy of 97%.

Although sign language recognition systems that use static images as inputs are simpler to train, movement is an important part of sign recognition. Passos et al. [12] developed a two-step method for gesture recognition of single sign using videos from three publicly available databases for Libras recognition: CEFET/RJ-Libras [13], MINDS-Libras [14], and LIBRAS-UFOP [15]. The first stage performed feature space mapping using body parts segmentation through a CNN and 2D motion encoding through Gait Energy Image (GEI). Then dimensionality reduction is applied to the feature space, testing four different techniques: Singular Value Decomposition (SVD), Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), and Locally Linear Embedding (LLE). In cases where the dataset is unbalanced, Synthetic Minority Oversampling Technique (SMOTE) was applied as a data augmentation method. The classification pipeline tested five classifiers: k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Random Forest, XGBoost, and LightGBM. A random hyperparameter search was used for each algorithm. The accuracy achieved by the system was 85.40% for CEFET/RJ-Libras, 84.66% for MINDS-Libras, and 64.91% for LIBRAS-UFOP.

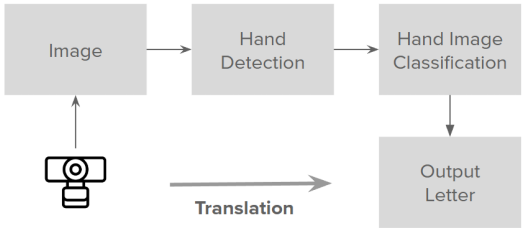

Figure 1. Proposed system for Libras alphabet sign recognition

3. Methodology

This work presents a simple version of a sign language recognition system that uses static images to classify letters of the Brazilian Sign Language (Libras) alphabet. As shown in Figure 1, the original objective was to build a model that, based on an image taken from a webcam, detects the hand region, extracts the region of the hand as a new image, and feeds it to a classifier that outputs the predicted letter.

Since the Libras alphabet includes letter signs that contain movement — H, J, K, X, and Z — those were excluded from the tasks, remaining 21 possible labels. For the hand detection task, a total of 3,150 images with 640x480 resolution were collected by a single person — 150 images for each letter — and the hand region was annotated. For each image, the annotated region was resized to 64x64 resolution to be the input in the classification task.

For the classification, the above-mentioned hand images were used along with a dataset of 46,401 images of hands performing the alphabet signs in Libras, with 64x64 resolution. The dataset was publicly available on Kaggle [16]. Therefore, the dataset used for the classification task had 49,551 images in total.

For the classification task, the dataset was divided into train, validation, and test in 60/20/20 proportion. The train and the validation sets were used for hyperparameter tuning in a CNN architecture. The number of convolution layers, the number of epochs, the learning rate, and the optimization function were the tested hyperparameters. After defining the best combination, the train and validation sets were fused into one and used to train the final model with the best hyperparameters. The final accuracy was evaluated in the test dataset. For the hand detection task, the experiment design is very similar, with the difference in the input images used. The experiments were conducted in Python 3.9.7 using PyTorch 1.11.0.

Architecture	MSE Validation
Features: Same as classification Epochs: 100 Learning rate: 0.01 Optimization: Adam	0,418767
Features: Same as classification Epochs: 100 Learning rate: 0.01 Optimization: SGD	0,082166
Features: ResNet50 Epochs: 100 Learning rate: 0.01 Optimization: SGD	0,000333
Features: ResNet50 Epochs: 100 Learning rate: 0.05 Optimization: SGD	0,000070
Features: ResNet50 Epochs: 200 Learning rate: 0.05 Optimization: SGD	0,000035
Features: ResNet50 Epochs: 200 Learning rate: 0.1 Optimization: SGD	0,000035
Features: ResNet50 Epochs: 400 Learning rate: 0.05 Optimization: SGD	0,000022

Table 1. Results of experiments for the hand detection model

4. Results

The final architecture for the classification model is shown in Figure 2. The model achieved an accuracy of 99.81% in the test set. Although this is a very high accuracy, the dataset used has a high level of similarity between images. For example, all images have a white background and the signs were performed by people with white skin. This can cause difficulties in real-life use of the system and overestimate the accuracy in these scenarios.

The hand detection task was more difficult as the number of input images was much smaller. A few different architectures were evaluated and the summary of results is shown in Table 1, including models with the same architecture of CNN layers as the classification model and models using the ResNet50 [8] CNN layers for the feature extraction step. The model with the best performance still didn't have a satisfactory result.

After the experiments were concluded, a Python program was used to allow users to consume the classification model in real-time. The program used webcam images to capture the letter signs. Since the hand detection model did not achieve a satisfactory result, it was not included in the final system. To capture only the hand region, a square was positioned at the top-right of the webcam image. The prediction is computed real-time and it is displayed above the square, as shown in Figure 3.

Although for the classification task the test set showed very high accuracy, letter signs that are similar to each other caused confusion in the final system. This is easily noticed by testing the classification model for some of these letters, as shown in Figure 4.

5. Conclusions and Future Work

This work presented a simplified system to translate letter signs from the Brazilian Sign Language (Libras) to the written letters. The CNN model for hand classification was trained from scratch and achieved a very high accuracy of 99.81%. However, Libras has some letter signs that are very similar to each other and can confuse the model. As for the hand detection, the number of images appeared not to be enough to train an efficient detector. Therefore, the detection step was removed from the final system. A Python program was developed to allow users to capture images and classify letter signs into text in real-time using a webcam.

A major limitation for the system is relative to the hand detection. An improvement could be made by acquiring more data to support the training process of the model. Also, the amount of similarities in the dataset used for training the classification model can lead to overestimating the accuracy of the system in real-life settings by it's lack of generalization in terms of background and skin tone, among other characteristics. Additionally, 5 letter sign were excluded from the tasks due to having movements. This could be solved by building a system that used short videos as input instead of static images.

Future work could be developed by exploring the limitations of this work cited above. Furthermore, the alphabet is only a small part of communication using Libras: most conversations use signs for whole words. Therefore, a system that supports communication in daily interactions must be able to recognize a single word or even a sequence of words that form a whole sentence.

References

[1] World federation of the deaf. https://wfdeaf.org/who-we-are/. Accessed: 2022-06-23. 1

[2] Pns 2019: país tem 17,3 milhões de pessoas com algum tipo de deficiência. https://censos.ibge.gov.br/2013-agencia-de-noticias/releases/31445-pns-2019-pais-tem-17-3-milhoes-de-pessoas-com-algum-tipo-de-deficiencia.html. Accessed: 2022-06-23. 1

[3] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, feb 2021. 1

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate

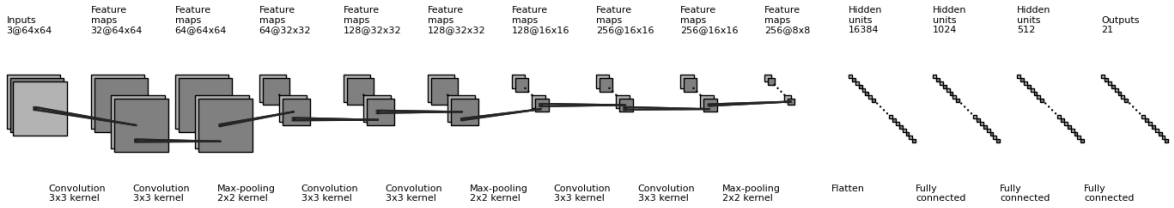

Figure 2. Final architecture for the classification model

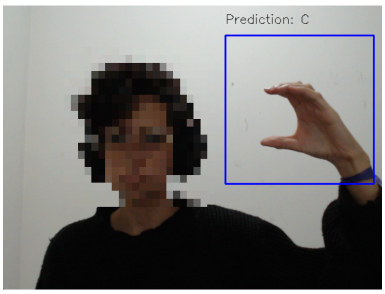

Figure 3. Example of use of the final program

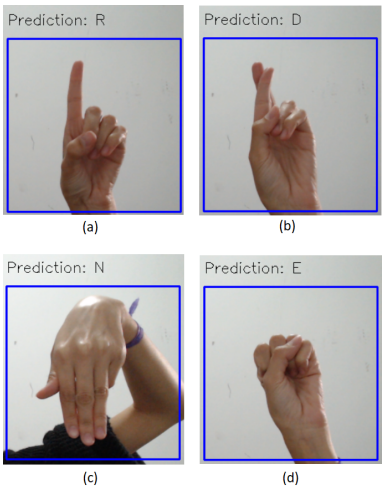

Figure 4. Examples of classification errors. (a) shows a D classified as a R, (b) shows a R classified as a D, (c) shows a M classified as a N, and (d) shows a S classified as an E.

object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, jan 2016. 1

[5] Ankita Wadhawan and Parteek Kumar. Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 32(12):7957–7968, jan 2020. 1, 2

[6] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 150:113336, jul 2020. 1

[7] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*, 79(31-32):22965–22987, jun 2020. 1, 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2, 3

[9] Chalearn lap isogd database. http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html. Accessed: 2022-06-24. 2

[10] Silas Luiz Furtado and Jauvane de Oliveira. Computer vision and neural networks for libras recognition. In *Anais do XVII Workshop de Visão Computacional (WVC 2021)*. Sociedade Brasileira de Computação - SBC, nov 2021. 2

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. 2

[12] Wesley L. Passos, Gabriel M. Araujo, Jonathan N. Gois, and Amaro A. de Lima. A gait energy image-based system for brazilian sign language recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(11):4761–4771, nov 2021. 2

[13] Priscila V. Gameiro, Wesley L. Passos, Gabriel M. Araujo, Amaro A. de Lima, Jonathan N. Gois, and Anna R. Corbo. A brazilian sign language video database for automatic recognition. In *2020 Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE)*. IEEE, nov 2020. 2

[14] Tamires Martins Rezende, Sílvia Grasiella Moreira Almeida, and Frederico Gadelha Guimarães. Development and validation of a brazilian sign language database for human gesture recognition. *Neural Computing and Applications*, 33(16):10449–10467, mar 2021. 2

[15] 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, apr 2021. 2

[16] Libras dataset - kaggle. https://www.kaggle.com/datasets/williansoliveira/libras. Accessed: 2022-04-22. 2