An approach for Brazilian Sign Language (BSL) recognition based on facial expression and k-NN classifier

Tamires Martins Rezende, Cristiano Leite de Castro
The Electrical Engineering Graduate Program
Federal University of Minas Gerais
Belo Horizonte, Brazil
Email: {tamires,crislcastro}@ufmg.br

Sílvia Grasiella M. Almeida
Department of Industrial Automation
Federal Institute of Minas Gerais - Ouro Preto
Ouro Preto, Brazil
Email: silvia.almeida@ifmg.edu.br

***Abstract*—The automatic recognition of facial expressions is a complex problem that requires the application of Computational Intelligence techniques such as pattern recognition. As shown in this work, this technique may be used to detect changes in physiognomy, thus making it possible to differentiate between signs in BSL (Brazilian Sign Language or LIBRAS in Portuguese). The methodology for automatic recognition in this study involved evaluating the facial expressions for 10 signs (to calm down, to accuse, to annihilate, to love, to gain weight, happiness, slim, lucky, surprise, and angry). Each sign was captured 10 times by an RGB-D sensor. The proposed recognition model was achieved through four steps: (i) detection and clipping of the region of interest (face), (ii) summarization of the video using the concept of maximized diversity, (iii) creation of the feature vector and (iv) sign classification via k-NN (k-Nearest Neighbors). An average accuracy of over 80% was achieved, revealing the potential of the proposed model.**

***Keywords*-RGB-D sensor; Brazilian Sign Language; k-NN; Facial expression.**

I. Introduction

Facial expressions are an important non-verbal form of communication, characterized by the contractions of facial muscles and resulting facial deformations [1]. They demonstrate feelings, emotions and desires without the need for words, and are an essential element in the composition of signs.

To determine the meaning of a sign – the smallest unit of the language – the location, orientation, configuration, and trajectory of both hands is essential. In addition to these characteristics, Non-Manual Expressions are features that can qualify a sign and add to its meaning, as well as being specific identifiers of a given sign [2].

The BSL recognition using computational methods is a challenge for a variety of reasons:

- There is currently no standardized database containing signs in a format that allows for the validation of computational classification systems;
- One sign is composed of various simultaneous elements;
- The language does not contain a consistent identifier for the start and end of a sign;
- Different people complete any given gesture differently.

To solve the first problem, a database was created with this study in mind. The following 10 signs - to calm down, to accuse, to annihilate, to love, to gain weight, happiness, slim, lucky, surprise, and angry - were chosen and captured 10 times, performed by the same speaker. An RGB-D sensor, the Kinect [3], operated through the nuiCaptureAnalyze [4] software was used.

According to [2], recent studies indicate the 5 main parameters of the BSL: point of articulation, hand configuration, movement, palm orientation and non-manual expressions. As the our focus is recognize facial expression, we chose 10 signs containing changes in facial expression during their execution. Then, this paper does an exploratory study of the peculiarities involved in non-manual sign language expression recognition.

Initially, a literature review of Computational Intelligence techniques applied to the sign recognition was accomplished. Promising results were reported recently in [2], in which feature extraction was done through the use a RGB-D sensor and a SVM (Support Vector Machine). However, the focus was in the motion of the hands. Another inspiring article is presented in [5]. While not addressing sign language, it applied Convolution Neural Networks method in the GAFFE dataset for facial expression recognition, achieving good results. Another important reference is [6], which proposed a system for recognizing expressions of anger, happiness, sadness, surprise, fear, disgust and neutrality, using the Viola-Jones algorithm to locate the face, extracting characteristics with the AM (Active Appearance Model) method, and classifying using k-NN and SVM.

Despite having distinct aims, these studies fit under the view of pattern recognition and served as the main references for the methodology proposed here.

It is important to note that the Maximum Diversity Problem was addressed in the Summarization. In the Classification step, cross-validation was used to identify the best value of k for k-Nearest Neighbor classifier and, through this, an average accuracy of 80% was reached.

The remainder of the paper is organized as follows: section II describes the database created for the validation of the method proposed. That is followed by an explanation of the methodology in section III. In section IV, the experiments and results are presented, with a conclusion in section V.

II. The Brazilian Sign Language database

For the creation of the database, the steps followed in [2] were used as reference. The first step was to choose the signs that contained changes in facial expression during execution (to calm down, to accuse, to annihilate, to love, to gain weight, happiness, slim, lucky, surprise and angry). After this, the signs were recorded. A scenario was created so that the speaker was in a fixed position (sitting in a chair), in approximately 1.2 meters from the RGB-D sensor, as shown in figure 1. This configuration was adopted since it allowed for focusing on the face. With the 10 signs each recorded 10 times with the same speaker, the balanced database had a total of 100 samples.

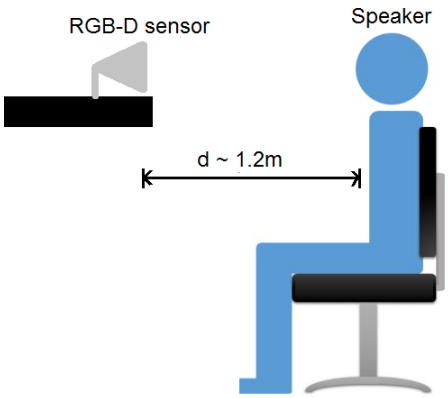


Fig. 1. Scenario created for recording the signs

Given the focus on facial expressions, nuiCaptureAnalyze was used to extract xy-coordinates of 121 points located across the face as in figure 2. These points served as the base descriptors for the face.

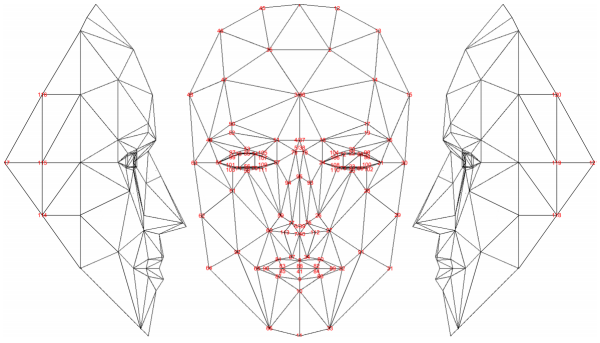


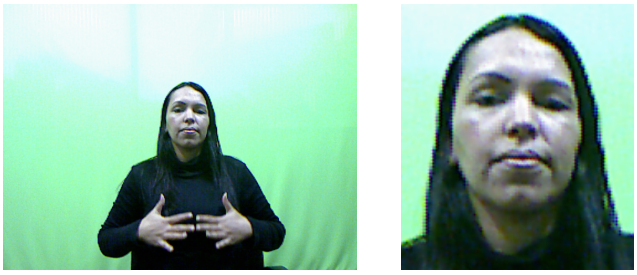
Fig. 2. Facial model used, with labeled points

III. Methodology

The following steps were defined so that the classification model was relevant to the available data extracted from the signs, with the objective of maximizing the model's accuracy. All the steps listed below were implemented in Matlab R2014a [7].

A. Detection of the region of interest

The original video contained view of the entire upper torso, so it was important to segregate the face specifically. This was done with the central pixel of the original frame as a reference, with the rectangular region cut out at a fixed coordinate. Figures 3a and 3b show a full frame and the separated area of interest respectively.



(a) Full frame (b) Selected face

Fig. 3. Facial selection

The face images are the inputs to the next step (Summarization) which will detect the most significant changes in facial expression.

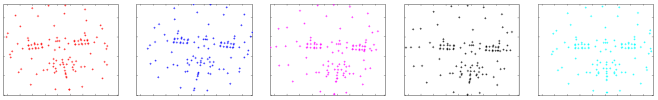
B. Summarization

This step was essential for the work, as it eliminates redundant frames allowing for a reduction of computational costs and more efficient feature extraction. Faced with the myriad of summarization techniques found in the literature, this study utilized the classic optimization problem, known as the Problem of Maximum Diversity [8], to extract the most relevant frames in the video.

After having selected the region of interest, the video recorded at approximately 30 fps was summarized through a process of selecting the n frames that contained the most diverse information. Based on the tests by [2], the specified value for n was five, such that each recording of each sign was summarized to a set of the five most significant frames as seen in figures 4 and 5. It is important to highlight that this summarization yielded feature vectors of equal size, regardless of the time that was taken to complete a given sign.

(a) 14 (b) 25 (c) 46 (d) 60 (e) 91

Fig. 4. The five most relevant frames extracted from a recording of the sign "to love"



(a) 14 (b) 25 (c) 46 (d) 60 (e) 91

Fig. 5. The 121 face points from the most relevant frames of the sign "to love"

C. Feature Vector

The objective of this step was to represent each sample in a manner that was robust and invariant to transformations. Each of the five frames returned from the summarization contained a descriptor formed from the concatenation of the xy-coordinates of the 121 facial points recorded by the nuiCaptureAnalyze software. Thus, the dimensions for a descriptor of a single frame are 1x242, with the following format:

$$D = \left[\begin{array}{cccccccc} x_1 & y_1 & x_2 & y_2 & \ldots & x_{121} & y_{121} \end{array} \right]_{1x242}$$

The feature vector for a sample is formed through the concatenation of the descriptors of its five selected frames. Thus, the final representation of any sample is a vector of length 1210.

$$Vector = \left[\begin{array}{ccccc} D_1 & D_2 & D_3 & D_4 & D_5 \end{array} \right]_{1x1210}$$

D. Classification

The k-NN method [9] was used for the classification step, as it is the recommended classifier for a database with few samples.

To determine the class of a sample m not belonging to the training set, the k-NN classifier looks for the k elements of the training set that are closest to m and assigns its class based on which class represents the majority of these selected k elements.

Initially, the 10 recordings for each of the 10 signs were randomized in order to prevent that the same recordings be selected for the training or testing groups. 80% of the data was selected for training, and 20% for testing, such that each train group had 8 samples and test group had 2 samples for each sign.

With the selected training data, a cross-validation was used to find the value for k that provided the highest accuracy rate, the k_{best}. Thus, the training data were divided into 5-folds of the same size and 5 cross-validation iterations were performed. For each one, 1-fold was removed for testing and the remaining were used for training, as shown in figure 6.

The equation 1 shows the range of k. Given there were 100 samples in total, the tested values for k were 1 to 10.

$$1 \leq k \leq \sqrt{number\ \ of\ \ samples} \tag{1}$$

For each value of k, 5 iterations of cross-validation were performed and the average accuracy was obtained. The value for k that provided the best result was used for the group test.

The average accuracy (acc_{avg}) and standard deviation (σ) for the testing set were obtained for the 10 iterations of the classification algorithm as shown in *Algorithm 1*. The metric distance used in k-NN method was the Euclidean distance.

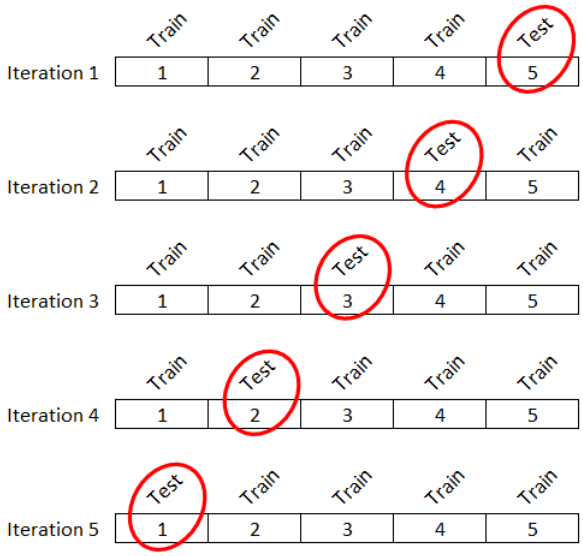


Fig. 6. Process for cross-validation using 5-folds

Algorithm 1: K-NN CLASSIFICATION

Input: Sign samples
Output: acc_{avg} and σ of the 10 iterations

Start
  for *w = 1 to 10* **do**
    Randomizes the samples of each sign
    train $\leftarrow$ 80% of the data
    test $\leftarrow$ 20% of the data
    for *k = 1 to 10* **do**
      $testV \leftarrow$ CROSS-VALIDATION(5-fold)
      $acc(k) \leftarrow$ K-NN$(testV, k)$
    end
    $[acc\ ind] \leftarrow max(acc)$
    $k_{best} \leftarrow ind$
    $acc_{test}(w) \leftarrow$ K-NN$(test, k_{best})$
  end
  $acc_{avg} \leftarrow mean(acc_{test}(w))$
  $\sigma \leftarrow std(acc_{test}(w))$
end
retorn acc_{avg}, σ

IV. Experiments and Results

It is known in literature that during the sign acquisition, distortions (offset, warping, etc.) can arise in different recordings of a same sign. This mainly occurs due to the natural displacement of the speaker's face during the sign recording. In order to overcome this problem and obtain samples invariant to distortions, some transformation procedures were applied on raw data. The experimental datasets considered in this study are described as follows.

First Experiment (EX.1): The first experiment consisted of the implementation of the methodology described throughout the paper, without any modification of the sign descriptors. In other words, the classification was performed with the raw data.

Second Experiment (EX.2): In the second experiment there was a processing of the database. For each recording, Z-Score Normalization was applied to all 5 frames of each point, based on equations 2 and 3.

$$x_{new} = \frac{x - \bar{x}}{\sigma(x)} \tag{2}$$

$$y_{new} = \frac{y - \bar{y}}{\sigma(y)} \tag{3}$$

Third Experiment (EX.3): In experiment 3, the data from each frame were updated according to equations 4 and 5. Using centroid normalization, each point was represented with reference to the mean point for that frame.

$$x_{newpointP} = x_{pointP} - \bar{x}_{frame} \tag{4}$$

$$y_{newpointP} = y_{pointP} - \bar{y}_{frame} \tag{5}$$

Table I contains the summary of the results from each of the experiments, as well as the values for k_{best} obtained from the 10 iterations of the algorithm. In figure 7, it is possible to compare the distribution of the percent accuracies for each of the three experiments.

TABLE I
RESULTS AFTER 10 EXECUTIONS OF THE CLASSIFICATION ALGORITHM.

Data	**Average Accuracy**	σ	k_{best}
EX.1	84%	8,76	1
EX.2	79%	6.99	1, 2 e 4
EX.3	83%	10.33	1

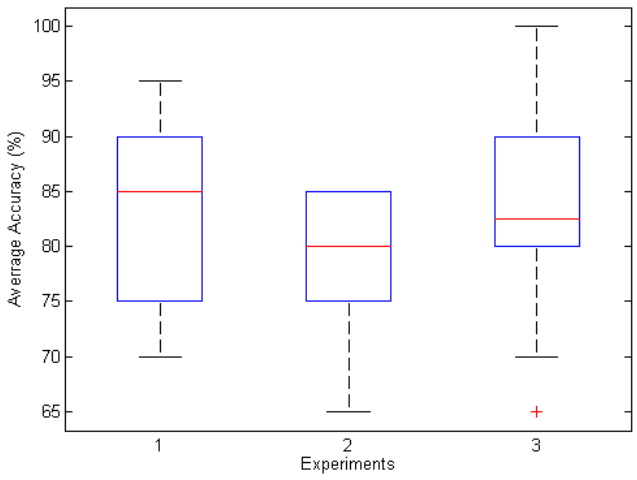


Fig. 7. Box plot of the percent accuracies for the three conducted classification rounds

V. CONCLUSIONS AND FUTURE WORK

The BSL recognition is a challenge problem. It is an area still in development that currently has no robust system for the classification of signs, such as phonemes, phrases or even a conversation, due to limitations resulting from the lack of a database with signs in BSL well structured. Despite these difficulties, working toward the development of a robust system is highly motivating, given the social impact that a system of this complexity can achieve.

In this paper, attention was taken to standardize the recording of all signs, randomize the sample to avoid bias in the data, in addition to following the statistical guidelines for choosing the best k.

With the aim of correctly classifying 10 signs, it was found that the methodology adopted had a considerable performance, achieving a maximum average accuracy of 84%. In addition, the system was shown to be robust when dealing with possible shifting of the face between different samples.

For future work, one of the intentions it to apply an SVM Multi-class classifier, adjusting the cost parameter C, that determines a balance between maximizing the margin and minimizing the misclassification [10], and the parameter γ, gamma of the kernel function. Another objectives are to verify the importance of the information about depth and to perform a selection of features reducing the dimensionality of the data.

ACKNOWLEDGMENT

The authors of this article would like to thanks PPGEE-UFMG for the incentive and guidance. The present work was completed with the financial support of CAPES - Brazil.

REFERENCES

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Patter Recognition 36*, pp. 259–275, 2002.
[2] S. G. M. Almeida, "Extração de características em reconhecimento de parâmetros fonológicos da língua brasileira de sinais utilizando sensores rgb-d," Ph.D. dissertation, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, 2014.
[3] Microsoft Windows, "Kinect." [Online]. Available: https://developer.microsoft.com/en-us/windows/kinect
[4] ——, "nuicaptureanalyse." [Online]. Available: http://nuicapture.com/download-trial/
[5] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1 – 8, 2015.
[6] F. J. C. Pedroso and E. O. Salles, "Reconhecimento de expressões faciais baseado em modelagem estatística," *XIX Congresso Brasileiro de Automática*, pp. 631–638, 2012.
[7] MathWorks, "Matlab r2014a."
[8] C. C. Kuo, F. Glover, and K. S. Dhir, "Analyzing and modeling the maximum diversity problem by zero-one programing," *Decision Sciences*, vol. 24, no. 6, pp. 1171–1185, 1993.
[9] E. Patrick and F. Fischer, "A generalized k-nearest neighbor rule," *Elsevier*, vol. 16, no. 2, pp. 128–152, 1970.
[10] M. H. Granzotto and L. C. Oliveira-Lopes, "Desenvolvimento de sistema de detecção de falhas baseado em aprendizado estatístico de máquinas de vetores de suporte," *XX Congresso Brasileiro de Engenharia Química*, pp. 1–10, 2014.