Using Neural Networks and 3D sensors data to model LIBRAS gestures recognition

Gabriel de Souza P. Moreira[1], Gustavo Ravanhani Matuck[1], Osamu Saotome[1], Adilson M. da Cunha[1]

[1] ITA – Brazilian Aeronautics Institute of Technology, São José dos Campos, SP, Brazil
gspmoreira@gmail.com, {gmatuck,osaotome,cunha}@ita.br

Abstract. This article applies neural networks models in the Brazilian Signs Language (*LInguagem BRAsileira de Sinais* - LIBRAS) alphabet recognition, composed by static and dynamic gestures. Gestures were recorded with a 3D camera sensor, providing coordinates from hands' fingertips. Deaf people, LIBRAS teachers, and students were involved in the recording process. The pre-processing involved frames sampling, normalization, and geometric transformations. For gestures recognition, artificial neural network models were trained and assessed to verify classification accuracy.

Keywords: Artificial Neural Networks, Sign Language Recognition, Gesture Recognition, LIBRAS, 3D Sensors.

1. INTRODUCTION

The Brazilian 2010 Census depicted 1,799,885 citizens in the country having great difficulty to hear. Signs languages are the natural languages of deaf communities. The Brazilian Signs Language (*Linguagem Brasileira de Sinais - LIBRAS*) is the sign language used by the majority of the deaf people in the Brazilian urban regions, and was standardized and recognized as the second official language by the Brazilian courts, on laws 10.436 (2002) and 5.626 (2005) [Decree 2005].

For many deaf-and-dumb people in Brazil, LIBRAS is the only way they can communicate. But, unfortunately, very few hearing people are knowledgeable and skilled on LIBRAS. That leads difficulties for deaf people to be assisted on hospitals, schools and universities, and government services.

This paper describes an investigation about the recognition of LIBRAS alphabet signs gestures, what would allow the spelling of Portuguese words. For the recognition, it was used a 3D sensor to capture hand and fingers coordinates, some pre-processing strategies, and neural networks to classify gestures.

Section 2 presents a brief literature research. Section 3 describes the conducted investigation, including data collection, pre-processing, learning models training and result analysis, followed by limitations and conclusions, in sections 4 and 5.

2. LITERATURE RESEARCH

Many attempts have been made to recognize signs languages gestures, generally represented by hands postures and gestures, and translate them to spoken languages letters, words, and expressions. The two major classes of hand tracking systems are: (1) data-glove-based, which requires electromechanical gloves with sensors; and (2) vision-based, which works with continuous image frames [Pistori and Neto 2004], usually regular RGB cameras or 3D sensors (like Microsoft Kinect [Kulshreshth et al. 2005]). For recognition models, some studies have been using machine-learning techniques, mostly Neural Networks and Hidden Markov Models (HMM).

In [Caputo et al. 2012], it is presented a 3D hand and gesture recognition system using Microsoft Kinect sensor, for dynamic gestures, and HD color sensor, for static gestures recognition. In [Kulshreshth et al. 2013], a real-time finger tracking technique was developed, using Kinect sensor signals, based on feature vectors calculation, using Fourier descriptors of equidistant points.

Researches addressing the Brazilian Signs Language have been conducted in the last years. In [Anjo et al. 2012], it is presented a real-time system to recognize a set of Libras alphabet (only static gestures), using MS Kinect and neural networks. In [Souza et al. 2007], authors use HMM to recognize 47 Libras gestures types, captured with regular RGB cameras.

In this article, we report the usage of vision based finger tracking, using Creative Senz3D GestureCam™ [Creative 2013], a set of pre-processing approaches, and neural network models to recognize all letters from Brazilian Signs Language (LIBRAS) alphabet, represented by static and dynamic gestures.

3. THE INVESTIGATION

The LIBRAS signs recognition process stages are Data Acquisition, Pre-processing, Models Training, and Results Analysis, as shown in Figure 1, as described in the following sections.

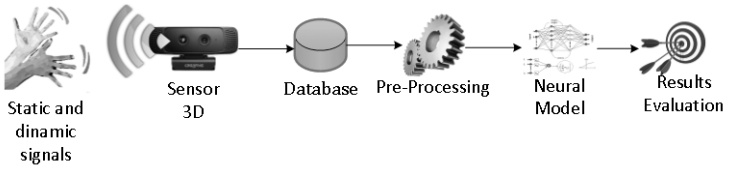

Fig. 1. The Investigation Process

3.1 LIBRAS alphabet

The LIBRAS alphabet letters are represented by 20 (twenty) static gestures and 6 (six) dynamic gestures (H, J, K, X, Y, Z) executed with movement. All alphabet signs are made with just one hand. This investigation considers only one-hand's data for recording and recognition.

3.2 Data Acquisition

In this stage, the gestures were recorded by using the COTS 3D sensor Creative Senz3D GestureCam™ [Creative 2013], and the Intel® Skeletal Hand Tracking Library (Experimental Release) [Intel 2013] for finger tracking, based on tracking approaches presented in [Melax et al. 2013]. In each frame, the library provided the full 6 Degrees Of Freedom (6-DOF) position and the orientation of 17 bones of a hand. The number of frames per second (fps) rate varied from 20-30 during recording.

Six (6) people volunteered to record LIBRAS alphabet signs, divided in three groups: deaf people (2), LIBRAS teachers (2), and students (2). For all of them, it was recorded two samples of all alphabets, except for one person, which had only one recorded sample. That resulted in 11 complete alphabet samples (268 letter gestures).

In this investigation, for each frame, it was recorded the absolute 3D coordinates (X, Y, and Z) of 5 fingertips and the center of the hand palm, based on the sensor distance, That resulted in 18 attributes with continuous position of hands and fingers.

3.3 Pre-processing

In order to prepare data for training the models, a pre-processing stage was conducted. The performed steps are described in the following list.

- Normalization to Hand Relative Coordinates – where the absolute fingertips coordinates were converted to relative coordinates, in relation to hand's center;
- Frames Sampling – where, for each gesture, it was recorded a different number of frames, because gesture time length varies, depending on alphabet letter and person style. The used sensors frames per second (fps) rate also varied, ranging from 20 to 30 fps. So, as the input for neural network model was fixed, it was necessary to normalize the input size, sampling a fixed number of frames for each gesture. This process selected 18 equidistant frames for each gesture, each frame containing all finger tips 3D coordinates, at some point in time; and
- Training Samples Transformations - In this investigation, there were only 11 samples for each alphabet letter. Therefore, a strategy was created and implemented to generate new samples based on training samples, by rotation and scaling geometric transformations.

3.4 Neural Network Model

The neural network used in this investigation was the traditional Multi-Layer Perceptron (MLP), with the back-propagation learning algorithm [Montini 2013]. All neurons used in the hidden and output layers had sigmoidal function activation. The Sum Squared Error (SSE) parameter was used to measure the learning performance. On this investigation, the MLP Network was composed by the input layer (where the processed data sensor were applied), one hidden layer (tested with 50 to 400 neurons, stepped by 50), and the output layer (with 26 neurons - the alphabet letters). Mathematically, only one hidden layer of neurons is enough to perform nonlinear mapping.

In order to provide a better way to train the neural network, a cross-validation process and a momentum parameter were applied. All data obtained from gestures signals were divided into training (64%), validation (18%), and test sets (18%). All range data used by the MLP Network is between [0,1]. Several simulations were performed with different MLP Neural Networks configurations.

3.5 The Results Analysis

After many simulations, the best MLP Network configuration correctly classified 61.54% (32/52) of unseen data during the network training (test set). Its architecture was composed of 200 neurons on the hidden layer.

This MLP Network training process was conducted for 1,000 epochs, after then the learning model started overfitting the sample data. Learning rate chosen by authors started with the value of 0.2, and decreased as epochs. The momentum parameter was used to improve the learning performance, avoiding local minima in the error surface.

Table 1 shows the classification results for the test set (two samples of each letter), presenting the MLP Neural Network difficulties to recognize some letters like "f", "k", "t", "x", and "z". On the other hand, the neural model correctly classified 10 letters with 100% of success ("a", "c", "d", "j", "l", "m", "o", "u", "v", and "w").

Table 1. The Alphabet Test Set classification analysis.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✗
✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗

4. LIMITATIONS

This research investigation presented some data acquisition quality limitations, which may have limited in some way the proposed neural network models accuracy. The chosen 3D sensor presented the following limitations: (1) The variability of frames per seconds rate may have introduced a bias when comparing samples, because of the temporality of gestures; and (2) An instability of finger tracking on experimental library (Intel® Skeletal Hand Tracking Library). Despite of the sensor, another investigation difficulty was to find people capable of accurately recording LIBRAS gestures, especially deaf people and teachers.

5. CONCLUSION AND NEXT STEPS

In this investigation, an initial research was described for recognizing the Brazilian signs language LIBRAS (*LInguagem BRAsileira de Sinais*) alphabet using Neural Networks over visual 3D sensor data. Six LIBRAS teachers, students, and deaf people have volunteered for recording the alphabet gestures. Some strategies were developed for pre-processing gestures data, to deal with temporality of gestures (frames sampling), normalizing coordinates, and increasing training samples, using geometric transformations (rotation and scaling).

Neural networks were assessed for gestures recognition, with different settings, training epochs, learning rates, and momentum factor. The best model correctly classified 61.53% patterns of the test set. Analyzing the results from this investigation, it can be observed that the MLP network was not as effective as expected when operating over noisy gestures data. Improvements in recognition models can be evaluated using more representative samples of gestures data. As a natural continuation of this research, still on its beginning phase, the authors intend to explore other 3D sensors, pre-processing approaches, and learning models, like Support Vector Machines and Hidden Markov Models.

REFERENCES

ANJO, M., PIZZOLATO, E., FEUERSTACK, S. A Real-Time System to Recognize Static Gestures of Brazilian Sign Language (Libras) alphabet using Kinect. *In Proc. of Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, Cuiabá, Brazil, 2012.

CAPUTO, M., DENKER, K., DUMS, B., and UMLAUF, G. 3D Hand Gesture Recognition Based on Sensor Fusion of Commodity Hardware. *Mensch & Computer*, 2012

CREATIVE. Creative Senz3D Gesture Camera - http://software.intel.com/sites/default/files/article/325946/creativelabs-camera-productbrief-final.pdf (2013)

DECREE. Brazilian Decree 5,626, December, 2005. http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2005/Decreto/D5626.htm

INTEL. Intel® Skeletal Hand Tracking Library SDK (Experimental Release) - http://software.intel.com/en-us/articles/the-intel-skeletal-hand-tracking-library-experimental-release (2013)

KULSHRESHTH, A., ZORN, C., and LAVIOLA, J. J.: Real-time Markerless Kinect based Finger Tracking and Hand Gesture Recognition for HCI. In *IEEE Symposium on 3D User Interfaces*, Orlando, USA, March 2013.

MELAX, S., KESELMAN, L., and ORSTEN, S. Dynamics based 3D skeleton tracking. In *Proc. Graphics Interface Conference, Saskatchewan*, Canada, May 2013.

MONTINI, D. Á.; MATUCK. G. R; DIAS, L. A. V., CUNHA, A. M., and RIBEIRO, A. L. P.: A Sampling Diagnostics Model for Neural System Training Optimization. In *10th ITNG 2013*. Las Vegas, USA, April, 2013

PISTORI, H. and NETO, J. An Experiment on Handshape Sign Recognition using Adaptive Technology: Preliminary Results. *Lect. Notes in Artificial Intelligence*, vol. 3171 (*XVII Braz. Symp. on Artificial Intellig.*), São Luis, Sept. 2004

SOUZA, K., DIAS J., PISTORI, H. Reconhecimento Automático de Gestos da Língua Brasileira de Sinais utilizando Visão Computacional. *III WVC - Workshop de Visão Computacional*, São José do Rio Preto, Brazil, Oct. 2007.