International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:12, No:12, 2018

A Motion Dictionary to Real-Time Recognition of Sign Language Alphabet Using Dynamic Time Warping and Artificial Neural Network

Marcio Leal, Marta Villamil

***Abstract*—Computacional recognition of sign languages aims to allow a greater social and digital inclusion of deaf people through interpretation of their language by computer. This article presents a model of recognition of two of global parameters from sign languages; hand configurations and hand movements. Hand motion is captured through an infrared technology and its joints are built into a virtual three-dimensional space. A Multilayer Perceptron Neural Network (MLP) was used to classify hand configurations and Dynamic Time Warping (DWT) recognizes hand motion. Beyond of the method of sign recognition, we provide a dataset of hand configurations and motion capture built with help of fluent professionals in sign languages. Despite this technology can be used to translate any sign from any signs dictionary, Brazilian Sign Language (Libras) was used as case study. Finally, the model presented in this paper achieved a recognition rate of 80.4%.

***Keywords*—Sign language recognition, computer vision, infrared, artificial neural network, dynamic time warping.

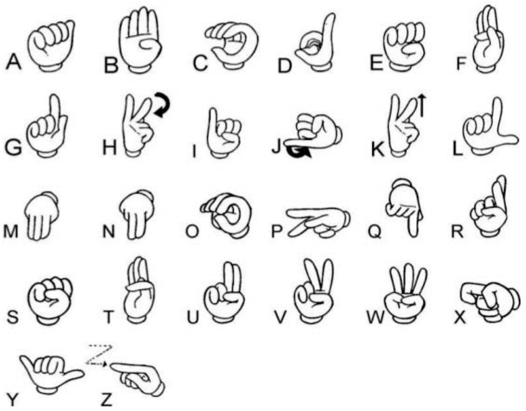

Fig. 1 Libras manual alphabet. Source: Silva et al. [8]

I. Introduction

According to World Health Organization, there are around 360 million deaf people in the world [1]. In despite of advances on development of valorization and inclusion of the deaf [2], [3], there is still difficulty for them to access basic services of health [4] and leisure [5].

Since inclusion of deaf in society still face the sign languages lack of knowledge of the listeners [6], computational tools that helps interaction deaf-listener can bring benefits to them [7]. Therefore the goal from this work is to present a motion dictionary from sign language manual alphabet. A secondary goal is to present a method to translate this manual alphabet into its written version. In view of the great range of existing sign languages, it was taken as a field for case study, the Brazilian Sign Language (Libras). Presented approach translate signs into text taking the recognition of two global parameters from sign languages: hand configurations and hand movements. Fig. 1 ilustrates both, hand configurations and movements that define Libras manual alphabet.

The approach has the following methodological steps. First, Leap Motion Controller (LMC)[1] capture positions of hand structure in real time. The virtual three-dimensional points are preprocessed and used as input parameters to a Multilayer Perceptron (MLP) Neural Network. Hand motion is classified by an algorithm called Dynamic Time Warping (DWT). Another parameters from sign languages, such as facial and body expressions are not considered in this work.

Therefore, the main contributions of this work are described as:

- **Libras signs dictionary**. Once that in the elaboration of this work Libras is used as a practical field, we are providing a dictionary composed of motion capture files of various signs acquired with help of specialists fluent in the language.
- **A new approach for sign languages recognition**. The approach from this work is based on the combination of motion capture by infrared technology in conjunction of MLP and the DTW algorithm, showing itself differentiated.

The remaining of the paper is organized as follow: Section II presents related work of sign language recognition using different sensors, with emphasis on those using LMC. Sections III-V presents the related technologies used on model. Method is presented in Section VI. Section VII presents results obtained with the method. And finally, Section VIII presents conclusions and future work.

II. Related Work

Nowadays, the works that proposed to recognize sign languages can be divided into those that use auxiliary devices, like electronic gloves, and those that use techniques of image processing and computational vision without the need to use

M. Leal and M. Villamil are with the Post Graduation Program on Applied Computing, Sinos River Valley University (UNISINOS), São Leopoldo, RS, 93.022-750 Brazil (e-mail: mr.leal@gmail.com, mbvillamil@unisinos.br).

[1] Motion capture dispositive explained in Section III

devices coupled to users body [9], [6]. In this session, we present works that use computer vision, giving emphasis to those that use the LMC.

One of the last works published about Libras recognition, Bastos et al. [9] proposed the use of descriptors to recognize 40 Libras signs through static images. In their method, at first, the authors use an MPNN to apply a binary mask in the images. The second step of the proposed method is the application of the Histogram of Oriented Gradients (HOG) and the Zernike Invariant Moment (ZIM) to extract characteristics to describe the images to finally classify the images through a second MPNN. This second Neural Network received as parameters the computed data in the previous steps. The proposed method achieved a recognition rate of 96.77%.

Other work that focused on image processing was Saha et al. [10], where the purpose was to recognize the 26 signs from American Sign Language (ASL) manual alphabet into static images through a new set of 11 features that feeded a MAdaline Neural Network. A preprocessing method was proposed to separate what is hand from the rest of the image, the preprocessing is done through the definition of a threshold within the 3 color spaces: RGB, HSV and YCbCr. According to the authors, the accuracy of correct answers was above 90%, what would justify the use of the proposed characteristics.

García-Bautista et al. used the RGB-D camera from Kinect to recognize 20 signs of the Mexican Sign Language (MSL) in [11]. The proposed method to classify the trajectory of the hand was the algorithm DWT. The method K-Fold Cross Validation was used for tests. The work showed an average hit of 98.57% for real-time recognition.

Pariwat and Seresangtakul [12] divided the signs from Thai Sign Language into 3 groups according to movements and combinations of characters. The work concentrated on identifying the group of static signs and involved the capture of 280x288 colored images from people wearing a black shirt on a blue background and an 8-stage process for hand segmentation. There are a total of 87 features collected from each image that are used in a Support Vector Machine (SVM) algorithm to classify the signs. At the end, an average precision of 91.2% was achieved.

The commercial version from LMC hardware is dated from 2014, when the first papers begin to be published.

One of the first papers using LMC [13], where Mohandes et al. did a comparison work using the MLP and the Nave Bayes Classifier. In this work, 12 of the features provided by LMC were selected, achieving an accuracy higher than 99% with the MPNN and 98% with the Bayesian Network to classify 28 signs of the static alphabet of the Arabic Sign Language (ArSL).

Chuan et al. [14] compared the K-Nearest Neighbor (KNN) and the Support Vector Machine (SVM) algorithms to classify 26 letters from ASL using LMC. Nine finger information was used to identify the signs, including the angles between each finger and the Z plane, distances between the finger joints and the center of the palm, among others. For the KNN classifier, was achieved 72.78% of accuracy and 79.83% to the SVM. Although the authors initially proposed to recognize 26 alphabet letters, there is no mention to J, the only letter from ASL which, invariably, needs movement to be recognized, once is the movement that differs it from letter I.

Another work that used LMC is from Elons et al. in [15] using a MLP to recognize 50 signs from ArSL. As test it was used 2 types of input data, distance between the position of each finger and LMC, and the distance between each finger and its subsequent. As a result, a maximum 88% accuracy is achieved for the second data type for static signs classification.

Fok et al. [16] used 2 LMCs and the Hidden Markov Model to classify numbers from 0 to 9 from ASL. As input data, it was used a series of information such as the orientation of each distal phalanx in relation to the direction of the palm, the distance from each finger tip to the central point of the palm, the distance of each finger tip and the tip of the subsequent finger, and finally the length of each finger. As final result, an average rate of 93.14% accuracy was achieved.

In 2016, Naglot and Kulkarni [17] used a MLP with LMC to recognize all the 26 signs the ASL alphabet. The work used as data the distance between fingertips and the center of the palm, beyond the distance between two consecutive fingers, reaching 96.15% of recognition rate.

The implemented system by Almasre and Al-Nuaim [18] used the LMC and Kinect to extract features. Kernel Support Vector Machine (KSVM) classified 28 static letters of the Arabic Sign Language. The system calculated 77 angles for the joints of the hands and 26 angles for every 2 bones, totaling 103 data. The use of PCA reduced the dimensionality of these data from 103 to 36, which represented 99% of variation. As result, an accuracy of 86% was presented for the set of data test.

The present work differs from others about the same theme on combining motion capture by infrared (IR) technology and the algorithms MLP and DTW.

As we are using Libras as practical field, we built a dictionary composed of motion capture files of various signs acquired with the help of specialists fluent in the language. This motion capture dictionary shows itself differentiated by not existing something similar for this specific language.

III. Leap Motion Controller

LMC is a small device that, through two IR cameras and three infrared emitters LEDs, can capture and rebuild the hand structure in real time into a three-dimensional cartesian coordinate environment centered on its top. The device has a capturing area around 61 cm from above its surface and its coordinate system is illustrated on Fig. 2.

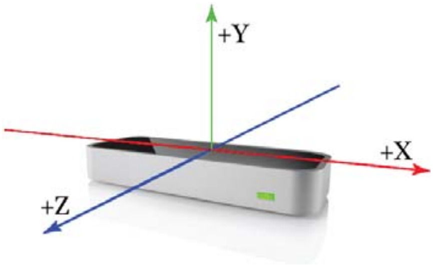

Fig. 2 O Leap Motion Controller (LMC). Source: Leap Motion [19]

The LMC programming interface provides types of data, some of them are described below:

- **Hand information:** the interface provides some information like which hand is being detected (left or right), position of the palm center, direction and speed of the palm. There is the information about the arm which the hand is associated;
- **Fingers information:** finger-related information such as direction, length, beginning and end position of each bone are some of the information provided.

Fig. 3 illustrates some information provided by the LMC programming interface.

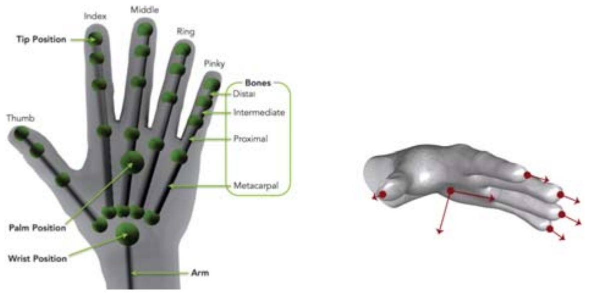

Fig. 3 The virtual hand hierarchy. Source: Adapted from Leap Motion [19]

LMC was chosen for its capability to work and deliver informations about the hand in real time.

IV. Multilayer Perceptron Neural Network

Artificial neuron is a fundamental unit from an artificial neural network. It is defined by three basic elements: a set of synapses, defined as a weight, more specifically, an input signal x_j connected to neuron k is multiplied by the weight of the synapse w_{kj}. The second element is an summing junction, responsible for adding the result of the multiplication of the input signals by the synapses of the neuron. Finally, the neuron has an activation function, which defines the amplitude from the output signal at a finite value [20]. An artificial neuron is illustrated by Fig. 4.

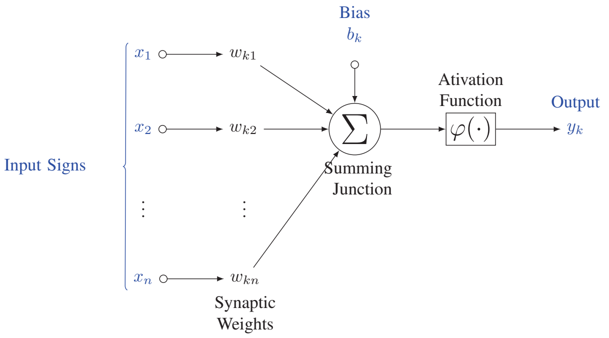

Fig. 4 The representation of an artificial neuron. Source: Haykin [20]

MLP is a type of artificial neural network used for classification problems and presents success even in the resolution of complex problems [20]. This kind of artificial neural network consists of a set of nodes (sensory units) arranged in layers. These layers are disposed in input layer, followed by intermediate (hidden) layers, which are followed by the output layer of the network.

Fig. 5 shows a typical MLP. Note that it has only 3 layers, among them 1 single hidden layer. The input signal is represented by the N-dimensional variable X_p, just as the output is represented by the M-dimensional variable Y_p. The variables W_{hi} and W_{oh} represents the matrices of weights that weighted the connections between neurons. In case of the W_{hi} matrix, this ponders the connections of the input layer neurons with the intermediate layer and presents a $N \times Nh$ dimension. Already the matrix W_{oh}, ponders the connections between the neurons of the middle layer with the output layer and presents a $Nh \times M$ dimension.

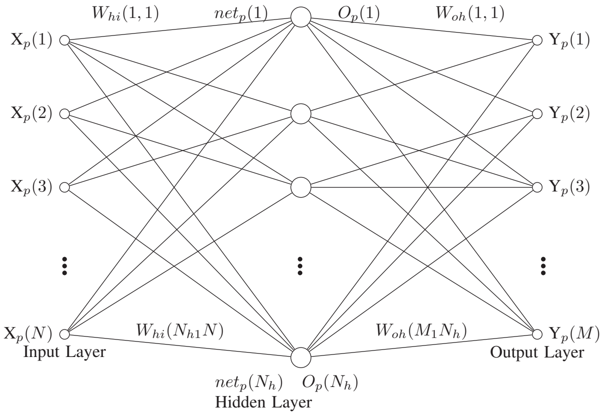

Fig. 5 Multilayer Perceptron Neural Network. Source: Walter and Michael [21]

The multi-layer perceptron neural network has been successfully applied to solve several complex problems, becoming a popular algorithm [20]. Once network is trained, it can be used to classify hand configurations in real hand in a very satisfactory way, therefore the algorithm was chosen for this purpose.

V. Dynamic Time Warping

Euclidean distance is the most established distance measure for comparison between time series, it measures dissimilarity between two series, but can be very sensitive to distortions occurring in the time axis [22]. Many applications require more flexibility in terms of observation.

In time series analysis, DTW is an algorithm to measure the similarity between two sequences of values that can vary in speed. For example, similarities in walking could be detected using DTW, even if one person was walking faster than other, or if there were accelerations and decelerations during an observation [23]. The DTW was applied to temporal sequences of audio [24] and video [25] data sequences, since data can be transformed into a linear sequence for analysis with DTW.

DTW is usually calculated using a dynamic programming algorithm. Equation (1) describes the initial condition of the algorithm.

Equation (2) presents the recurrence relation of the DTW algorithm, where $i = 1 \dots N$ and $j = 1 \dots M$. $c(x_i, y_j)$ is the cost of combining two observations x_i and y_j. The resulting value in $dtw(N, M)$ is the DTW distance between x and y. Thus, the algorithm fills, iteratively, a matrix with the lowest cumulative cost for all alignments to each pair of observations to be combined.

$$dtw(i,j) = c(x_i, y_j) + min \begin{cases} dtw(i-1,j) \\ dtw(i,j-1) \\ dtw(i-1,j-1) \end{cases} \tag{1}$$

Finally, Fig. 6 shows an example of the optimal nonlinear alignment found by the algorithm and how it is represented in the resulting matrix.

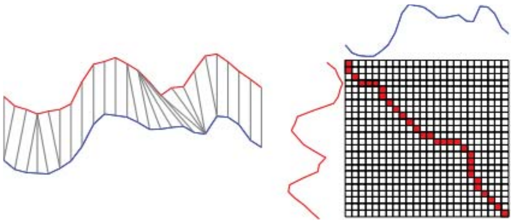

Fig. 6 Optimal nonlinear alignment (left) and the resulting matrix obtained by the dynamic programming algorithm, highlighting the ideal alignment (right). Source: Silva and Batista [22]

VI. Method

Our method identifies letters from alphabets of sign languages, taking the letters A to Z from Libras as pratical field, and it follows 4 steps, which are illustrated by Fig. 7 and are described below.

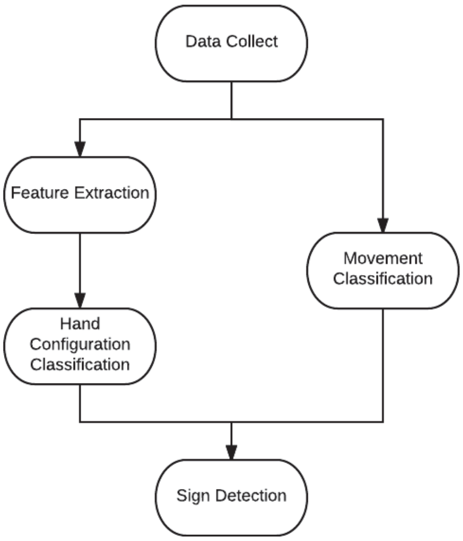

Fig. 7 Data flow fluxogram

a) Data collect: The dataset is collected using LMC and Fig. 3 illustrates some of the information delivered by the device that are: position of tip from each finger, position of each hand joint, position of palm center, wrist position, hand direction, hand normal and direction of each finger. Some adicional information like elbow position and hand speed are provided. A total of 104 data from LMC are used to compose the dataset.

All information formed the dataset files consiting of 750 samples, 30 of each letter we want to to recognize. 500 of them are used for training the neural network and 250 are used for testing.

b) Pre-Processing: In terms of pre-processing, based on article [17] from Naglot and Kulkarni, information about the tip from each finger and the center of the palm was selected. After capturing these data, the euclidean distance between each fingertip and the center of the palm is calculated, generating a set of 5 data that are used to train a MLP. A set of 5 more data is calculated refering to the euclidean distance from each fingertip to its subsequent, generating a total set of 10 data.

c) Hand configuration classifying: In Libras, there are 20 possible hand configurations identified in the 26 manual alphabet letters. For manual alphabet classifying, a MLP with three layers and backpropagation as training method was developed. As we have 10 pre-calculated input data, an input layer with 10 neurons, one for each calculated data at pre-processing stage was used. The output layer consists of 20 neurons since it is necessary 20 hand configurations to perform the whole Libras manual alphabet. Fig. 8 illustrates the implemented neural network.

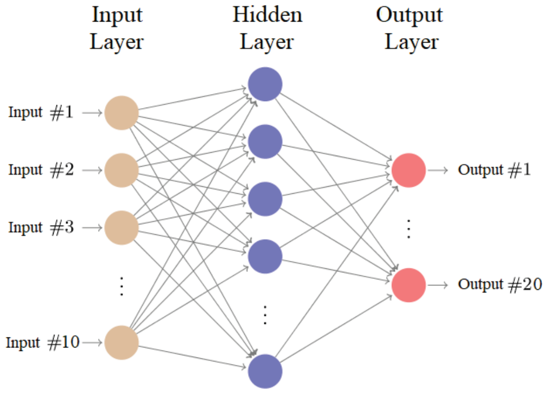

Fig. 8 MLP hand configuration classifyer

The smallest number of hand configurations than letters on manual alphabet happens because some of the letters use the same hand configuration, therefore, in a first moment, we can devide letters into six groups taking into account use of the same hand configurarion, such division can be verified at Fig. 9.

d) Motion classifying: The next step is to identify the hand motion. Into the manual alphabet there are 4 diferent kinds of motion: hand stopped, z motion, up motion and half circle. To classify this different movements, the algorithm called Dynamic Time Warping (DWT) that generates a distance from real time detection to each capture from data base, the shortest distance characterizes the recognition.

Once the hand configuration is detected, motion is used to classify letters into the groups 1, 3 and 4. To the other groups, motion is ignored since into groups 2 and 5 the differece relies only where hand is pointing and, into group

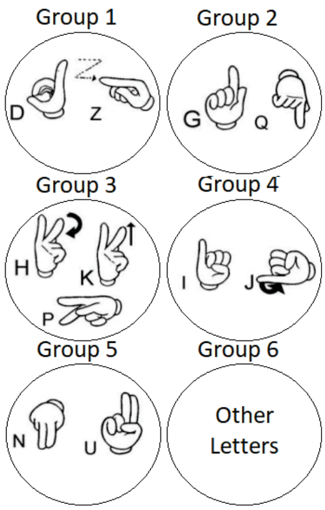

Fig. 9 Manual alphabet groups. Source: Adapted from Silva et al. [8]

6, is not necessary to make any other distinction beyond hand configuration detection.

e) Sign Detection: To join hand configuration and hand motion detections, a state machine was developed.

As showed on Fig. 10, states G1 to G6 stands to recognize hand configurations frame by frame, while states D, Z, G, Q, H, K, P, I, J, N and U are final states to recognize letters with motion.

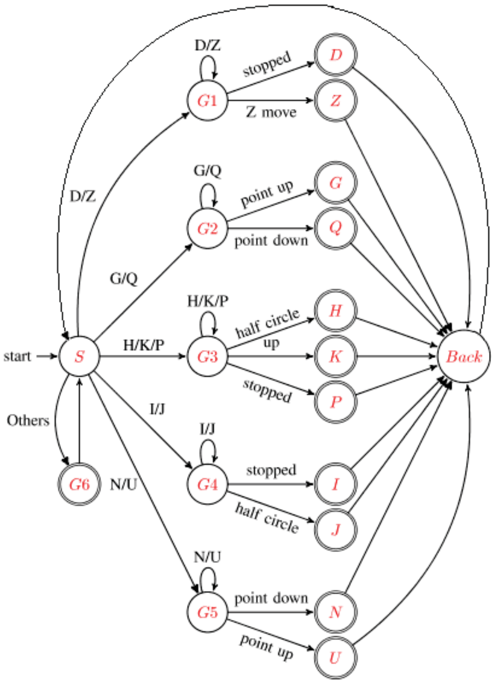

Fig. 10 The developed state machine

VII. Results

This article presents a model of recognition of two global parameters from sign languages: hand configurations and hand motion.

After the capture of hand structure and its joints into a virtual three-dimensional space, was used a MLP with backpropagation as training method to classify hand configurations, then DTW algorithm was used to classify motion.

Once the hand configurations and its motion are properly identified, the Table I presents the recognition rates for each letter from Libras alphabet, used as a case study. The main information that can be taken from referenced table is that the developed model can work with an accuracy of 80.4%.

TABLE I
Classification Rates

Word Recognized	Recognition Rates
A	90.0%
B	90.0%
C	80.0%
D	80.0%
E	90.0%
F	80.0%
G/Q	90.0%
H	70.0%
I	70.0%
J	70.0%
K	70.0%
L	90.0%
M	80.0%
N	70.0%
O	80.0%
P	60.0%
R	60.0%
S	90.0%
T	90.0%
U	80.0%
V	90.0%
W	90.0%
X	80.0%
Y	90.0%
Z	80.0%
Total	80.4%

VIII. Conclusion and Future Work

Capturing hand structure for sign language recognition requires a certain level of accuracy, mainly due to some similarities that may affect the accuracy of the recognition, such as hand configurations from letters O and C, and the letters K, H and P, for example, which differ only by its motion. Therefore, as future work, we will to work with 2 LMCs to try to increase the accuracy from motion capture by avoiding finger occlusions.

For this paper, we worked with sign languages alphabet recognition. As could be seen, letters can be computacionally identified by hand configuration and its motion. Signs with a more complex meaning than just alphabet, the ones similar to words from spoken languages, can be identified in a similar way, once the alphabet use a subgroup of this hand configurations and motion. Therefore, still as future work, we intend to expand the range of recognized signs.

One more important point is that, as we developed a model to computacionally recognize sign languages, taking Libras as study case, this work can be extended and replied to other sign languages.

References

[1] World Health Organization, "Deafness and hearing loss," 2015, available at: <http://www.who.int/mediacentre/factsheets/fs300/en/>. Accessed in: May 18, 2017.

[2] B. S. Guedes, "Sobre surdos, bocas e mãos: saberes que constituem o currículo de fonoaudiologia," Master's thesis, Universidade do Vale do Rio do Sinos, 2010.

[3] P. H. Witchs, "A educação de surdos no estado novo: práticas que constituem uma brasilidade surda," Master's thesis, Universidade do Vale do Rio dos Sinos, 2014.

[4] M. T. de Souza and R. Porrozzi, "Ensino de libras para os profissionais de saúde: uma necessidade premente," *Revista Práxis*, vol. 1, no. 2, 2009.

[5] I. S. X. de França, J. da Silva Aragão, A. S. Coura, C. E. N. K. Vieira, J. F. da Silva, and G. K. P. Cruz, "A relação entre atividades de lazer e níveis glicêmicos de adultos surdos," *Northeast Network Nursing Journal*, vol. 14, no. 6, 2013.

[6] E. M. Flores, J. L. V. Barbosa, and S. J. Rigo, "Um estudo de técnicas aplicadas ao reconhecimento da língua de sinais: novas possibilidades de inclusão digital," *RENOTE*, vol. 10, no. 3, 2012.

[7] P. R. Rodrigues and L. R. G. Alves, "Criar e compartilhar games: novas possibilidades de letramento digital para crianças surdas," *RENOTE*, vol. 12, no. 2, 2014.

[8] F. I. da Silva, F. Reis, P. R. Gauto, S. G. de Lima da Silva, and U. Paterno, *Aprendendo Libras como segunda lngua: nvel bsico*. Federal Institе from Santa Catarina, 2017.

[9] I. L. Bastos, M. F. Angelo, and A. C. Loula, "Recognition of static gestures applied to brazilian sign language (libras)," in *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*. IEEE, 2015, pp. 305–312.

[10] S. Saha, R. Lahiri, A. Konar, and A. K. Nagar, "A novel approach to american sign language recognition using madaline neural network," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2016, pp. 1–6.

[11] G. Garca-Bautista, F. Trujillo-Romero, and S. O. Caballero-Morales, "Mexican sign language recognition using kinect and data time warping algorithm," in *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Feb 2017, pp. 1–5.

[12] T. Pariwat and P. Seresangtakul, "Thai finger-spelling sign language recognition using global and local features with svm," in *2017 9th International Conference on Knowledge and Smart Technology (KST)*, Feb 2017, pp. 116–120.

[13] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in *2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, June 2014, pp. 960–965.

[14] C. H. Chuan, E. Regina, and C. Guardino, "American sign language recognition using leap motion sensor," in *2014 13th International Conference on Machine Learning and Applications*, Dec 2014, pp. 541–544.

[15] A. S. Elons, M. Ahmed, H. Shedid, and M. F. Tolba, "Arabic sign language recognition using leap motion sensor," in *2014 9th International Conference on Computer Engineering Systems (ICCES)*, Dec 2014, pp. 368–373.

[16] K. Y. Fok, N. Ganganath, C. T. Cheng, and C. K. Tse, "A real-time asl recognition system using leap motion sensors," in *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Sept 2015, pp. 411–414.

[17] D. Naglot and M. Kulkarni, "Real time sign language recognition using the leap motion controller," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 3, Aug 2016, pp. 1–5.

[18] M. A. Almasre and H. Al-Nuaim, "Recognizing arabic sign language gestures using depth sensors and a ksvm classifier," in *Computer Science and Electronic Engineering (CEEC)*, Sept 2016, pp. 146–151.

[19] Leap Motion, 2014, available at: <https://developer.leapmotion.com/>. Acessed in: May 9, 2017.

[20] S. O. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Pearson, 2008.

[21] H. Walter and T. Michael, "Recent developments in multilayer perceptron neural networks," in *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference*, 2005.

[22] D. F. Silva and G. E. Batista, "Speeding up all-pairwise dynamic time warping matrix calculation," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 837–845.

[23] T. Prtzlich, J. Driedger, and M. Mller, "Memory-restricted multiscale dynamic time warping," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 569–573.

[24] M. Muller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization." in *ISMIR*, 2006, pp. 192–197.

[25] A. Gupta, J. He, J. Martinez, J. J. Little, and R. J. Woodham, "Efficient video-based retrieval of human motion with flexible alignment," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.