Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 5, October 2020, pp. 2082~2089
ISSN: 2302-9285, DOI: 10.11591/eei.v9i5.2440

A gesture recognition system for the Colombian sign language based on convolutional neural networks

Fredy H. Martínez S.[1], Faiber Robayo Betancourt[2], Mario Arbulú[3]
[1]Universidad Distrital Francisco José de Caldas, Facultad Tecnológica, Colombia
[2]Universidad Surcolombiana, Facultad de Ingeniería, Departamento de Electrónica, Colombia
[3]Universidad Nacional Abierta y a Distancia (UNAD), Colombia

Article Info

Article history:

Received Jan 20, 2020
Revised Mar 6, 2020
Accepted Apr 4, 2020

Keywords:

Convolutional network
Deaf-mutes people
Nasnet
Sign language
Spanish text
Visual techniques

ABSTRACT

Sign languages (or signed languages) are languages that use visual techniques, primarily with the hands, to transmit information and enable communication with deaf-mutes people. This language is traditionally only learned by people with this limitation, which is why communication between deaf and non-deaf people is difficult. To solve this problem we propose an autonomous model based on convolutional networks to translate the Colombian Sign Language (CSL) into normal Spanish text. The scheme uses characteristic images of each static sign of the language within a base of 24000 images (1000 images per category, with 24 categories) to train a deep convolutional network of the NASNet type (Neural Architecture Search Network). The images in each category were taken from different people with positional variations to cover any angle of view. The performance evaluation showed that the system is capable of recognizing all 24 signs used with an 88% recognition rate.

Corresponding Author:

Fredy H. Martínez S,
Universidad Distrital Francisco José de Caldas,
Cl. 68d Bis ASur #49F-70, Bogotá, Colombia.
Email: fhmartinezs@udistrital.edu.co

1. INTRODUCTION

The term deaf-mute is used to describe people who are deaf from birth and who have great difficulty communicating by voice [1, 2]. Formerly it was thought that deaf-mute people could not communicate, however, this is no longer correct, many visual sign languages have developed, both written and spoken [3]. Many countries have their own sign language which is learned by their deaf community [4, 5]. The problem is that normally this language is unknown to non-deaf people, which makes communication outside the group impossible. However, against the current precepts of equality and social inclusion, it is clear that these sign languages are as important as the spoken language [6, 7]. It is estimated that 5% of the world's population has disabling hearing loss, in Colombia by 2019 this corresponds to just over 2.4 million people [8].

Sign language is a language formed by gestures that can be visually identified [9]. Normally, postures of the hands and fingers, and even facial expressions are used, all of them with a certain meaning to form symbols according to a code [10]. Each sign language is made up of tens or even hundreds of different signs, and in many cases, the differences between them are just small changes in the hand. In addition, like written language, each person prints his or her personal mark on the representation of these signs [9]. Sign language recognition systems develop schemes and algorithms that allow these symbols to be correctly identified and translated into text recognizable by the community at large. There are two strategies for the development of these systems, schemes based on artificial vision and schemes based on sensors.

In the first case, these are artificial systems that use digital processing from images (or video frames) captured from digital cameras [11, 12]. The algorithms usually use filters and digital image processing, or pattern recognition schemes using bio-inspired schemes [13-15]. In the second case, sensors are used to detect specific elements in the hands such as gloves, colors, flex sensors or textures [16, 17]. This scheme requires the modification of normal language communication conditions, which is why it is a less attractive solution.

Among the strategies implemented in these recognition systems, those with the highest performance and impact on development and research should be highlighted, particularly those that can be used without modifying the natural behavior of the person. A widely used strategy both for the case of digital image processing (which is where it is, in fact, most used) and for the case of detection of specific sensors is the segmentation of the signal's field information (hand), and subsequent extraction of characteristics from each segment for subsequent classification [11, 18]. One of the most widely used tools for the classification process is the neural networks, and more recently, the deep learning [19, 20]. These tools have the advantage of being able to generalize specific characteristics of the signs and detect them in different hands and people, in a similar way as the human brain does [21].

It should also be noted that there are three translation strategies that can be performed by a sign language translation system, these are spelling (letter to letter), single words and continuous sentences separated by signs [16]. Fingerspelling is used in situations where new words, names of people and places, or words without known signs are spelled in the air by hand movement [22, 23]. This is the most basic scheme, and where we find the most research. Isolated word recognition analyzes a sequence of images or hand movement signals that ultimately represent a complete signal [24, 25]. In this scheme, letter recognition should be supported by a written language dictionary to identify complete words. Finally, the most complex scheme is that of continuous sentences, which integrates grammatical strategies, and which is, in fact, the most important for real-life communication situations [26, 27].

This article proposes a strategy for the recognition of static symbols of the Colombian Sign Language (CSL) for their translation into Spanish letters (fingerspelling). The system uses a deep neural network trained and tuned with a database of 24000 images distributed along 24 categories. The performance of the scheme was measured using traditional machine learning metrics. The rest of this article is structured as follows. Section 2 presents preliminary concepts and problem formulation. Section 3 illustrates the design profile and development methodology. Section 4 we present the preliminary results. And finally, in Section 5, we present our conclusions.

2. PROBLEM FORMULATION

For some years now, the research group has been carrying out a research project aimed at developing a robotic platform (assistive robot) for the care of children, sick people and the elderly. Within this project, specific works are framed on path planning in observable dynamic environments, direct human-machine interaction, automatic identification of emotions in humans, and the problem here posed of identification of visual signs corresponding to the CSL. All these problems separately correspond to open research problems in robotics, which the research group tries to solve to implement complex tasks in service robotics.

In the specific case of the visual sign recognition system, the aim is to adapt the robotic platform for direct interaction with deaf-mute people. The robot has two cameras in the head to capture images with signs and has a touch screen to display information in response to the user. Therefore, the proposed identification system should use as an input mechanism the digital image of the sign represented by the user, and as verification and response mechanism the representation of the identified letter on the screen. The idea is to develop a recognition method that can be scaled by increasing the size of the vocabulary, and that would be robust to the variations of the angle of image capture, and those due to the specific characteristics of the people (type of hand, skin color, clothing, height or variability in the position of the fingers). Different people have different hands with different dimensions and geometry, for this reason, the same signal made by two different people can produce different measurements when analyzing landmarks in images. These are variations that should not alter the operation of the classifier.

To represent a letter on the CSL, one of the hands is used (the use of the right hand is assumed, but the scheme can well be trained to identify either hand, insensitivity to the dominant hand) and all its fingers. Most signs correspond to static positions of the fingers, however, some letters require movement of the fingers and hand (letters Ñ, S, and Z). For this first model we will use only static positions of the hand, reason why these three letters are excluded, which implies that the classifier has a total of 24 categories as shown in Figure 1.

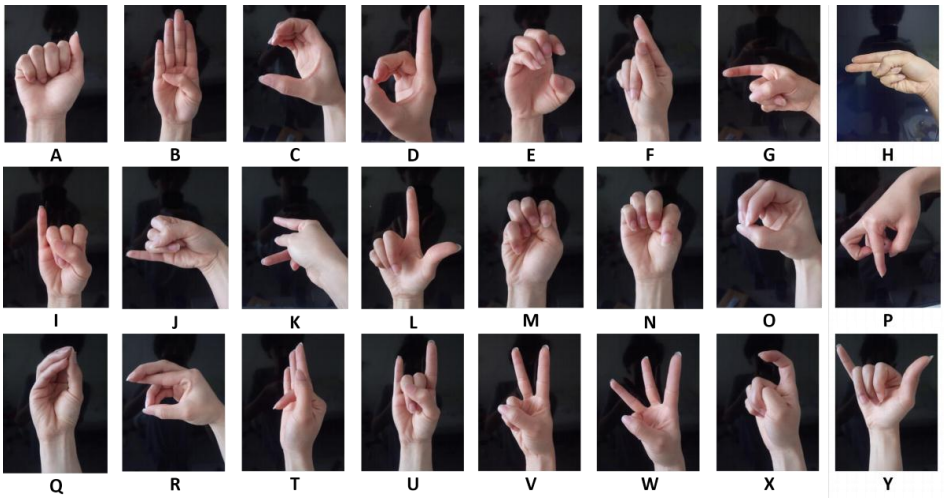

Figure 1. CSL alphabet used in the proposed recognition system

The robot must be in front of the user to capture the image corresponding to the sign, but there is not a correct fixed distance for the capture. Many sign language identification systems are distance-dependent, so the same sign represented by the same person but captured at different distances can produce different results. This is a problem for real systems since the sign is not distance-dependent from the user's perspective. However, the distance to the robot (and therefore to the image capture camera) must be consistent with the distance between two people talking, i.e. between about 0.2 to 0.7 m. The background of the image is also not restricted to a specific type or color since in real operation of the system the user can wear any clothing and be located in any space of the house, school or place of interaction with the robot. The system must be able to identify the parts of the foreground and the information encoded there.

3. RESEARCH METHOD

To implement the identification algorithm we use the NASNet (Neural Architecture Search Network) deep neural network. This convolutional network was introduced in early 2018 by the Google Brain team, and in its design, they sought to define a building block with high performance in the categorization of in a small set of images (CIFAR-10) and then generalized the block to a larger data set (ImageNet). In this way, this architecture achieves a high classification capacity and a reduced number of parameters as shown in Figure 2.

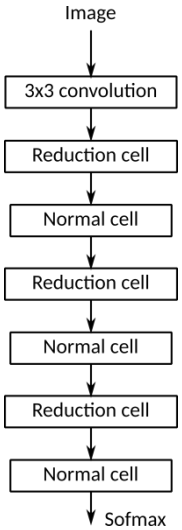

Figure 2. ImageNet architecture (NASNet)

We built our training database that can be seen in Figure 3. Our dataset was built by capturing images for each of the 24 categories in Figure 1. In total, we used 1000 images in each category for a total of 24000 images. The same number of images in each category was maintained to achieve a balance of classes in the trained model. To separate the images within each category, the images corresponding to each class were placed together in a folder labeled with the letter c, an order number, and the corresponding letter. Different people were used to capturing the images, in different locations, viewing angles, distances to the camera and lighting conditions. This was done to try to make the classification algorithm identify the information of common interest in each category, and to increase the robustness against changes in the environment.

To improve the performance of the network we randomly mix the images during training. With the same intention, we scale all images to a size of 256x256 pixels. In the model we have not considered the aspect ratio of the images, we believe that the information in the images is not altered by this change, besides, the image size is important for the training of the network. For training purposes we normalize the color value of each RGB matrix of the images to the range of 0 to 1, this corresponds to the operating value of the neural network. The database was randomly divided into two groups, one for training and one for testing. We used 75% of the data for training and 25% for performance evaluation. In the network design, the number of nodes in the input layer is defined according to the size at which the images are resized, i.e. 255x255x3 considering the three-color matrices. The number of nodes in the output layer is defined by the number of categories for the classification of the images, i.e. 24 nodes. As optimization and cost metrics in training, we use stochastic gradient descent, categorical cross-entropy, accuracy, and mse.

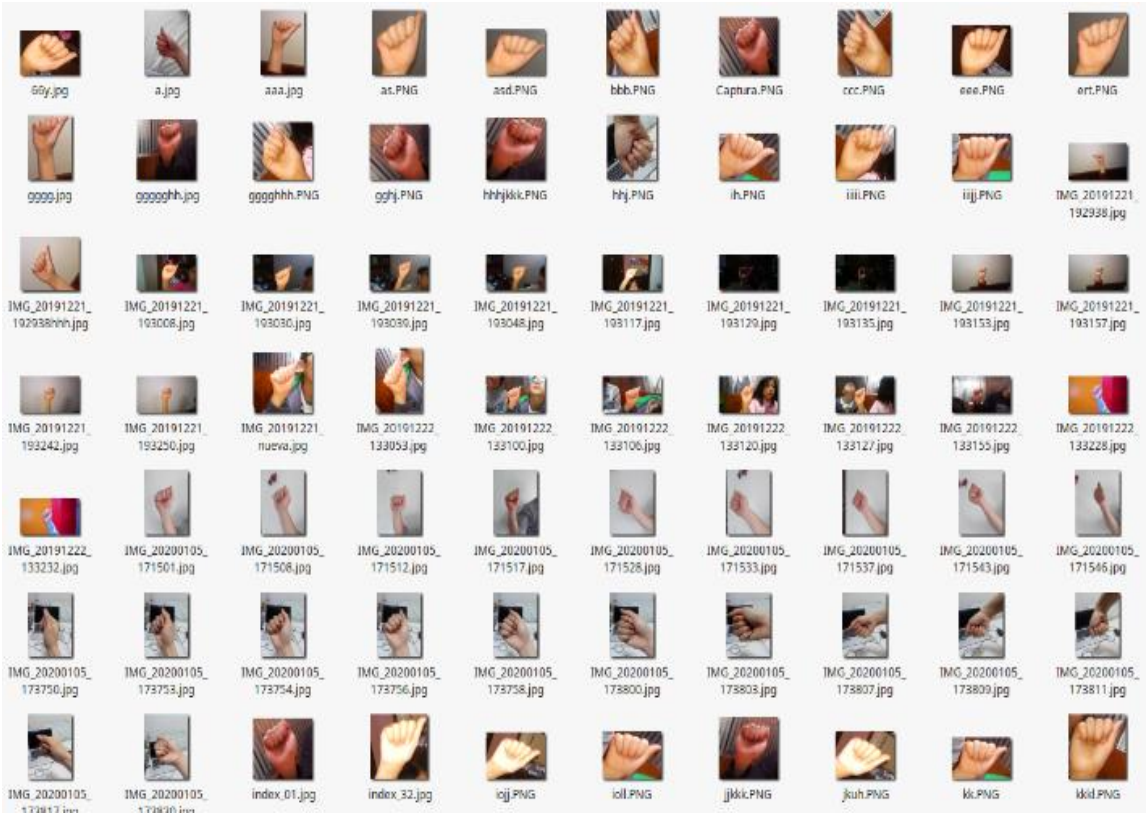

Figure 3. Sample database of CSL images used for model training

The neural network had a total of 4,295,084 parameters, of which a total of 4,258,346 were adjusted (trained). The neural model training code was developed in Python 3.7.3 within the Keras framework (using TensorFlow backend). As libraries, we use Scikit Learn 0.20.3, Pandas 0.24.2, Nltk 3.4, Numpy 1.16.2, and Scipy 1.2.1.

4. RESULTS AND DISCUSSION

The training was performed during 10 epochs, and as metrics of adjustment during the training, the function of loss of cross-entropy (categorical cross-entropy), the accuracy (or success rate) and the MSE (Mean Squared Error) were calculated. The results during the training of these three metrics, for both training and validation data, are shown in Figures 4, 5 and 6. As can be seen, the reduction of error is constant for both training and validation data, with equivalent behavior for accuracy, which follows correct learning without problems of over-or under-adjustment.

After the model was trained, its performance was checked against the validation data by means of its Confusion Matrix, from the Precision, Recall and F1-score metrics, and by means of the ROC (Receiver Operator Characteristic) curve. Figures 7, 8 and 9 show the results achieved. The diagonal in Figure 7 shows a very good classification hot zone, which coincides with the data by category of the metrics summarized in Figure 8.

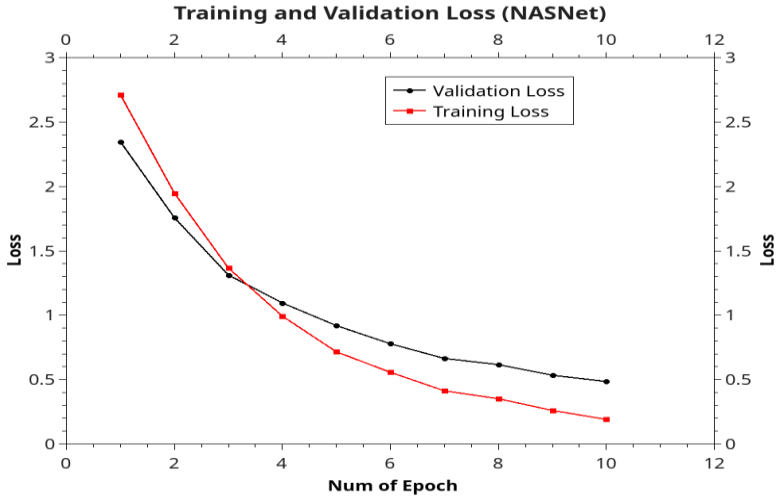

Figure 4. Training and validation loss

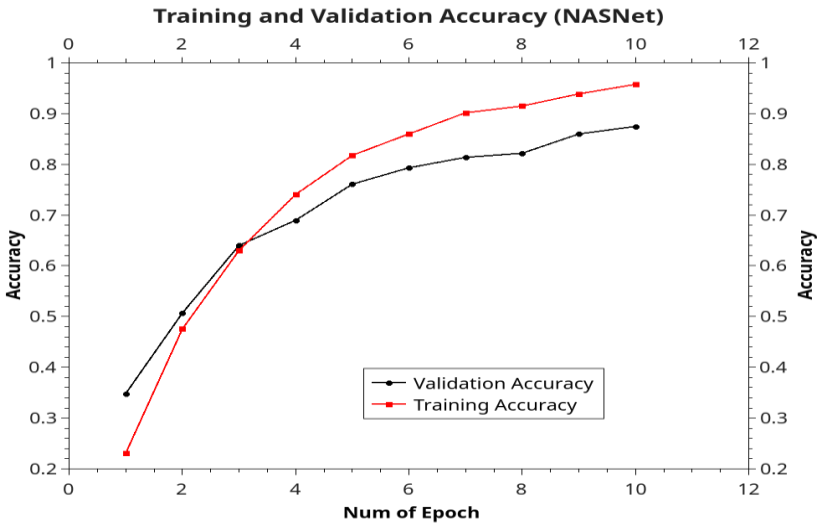

Figure 5. Training and validation accuracy

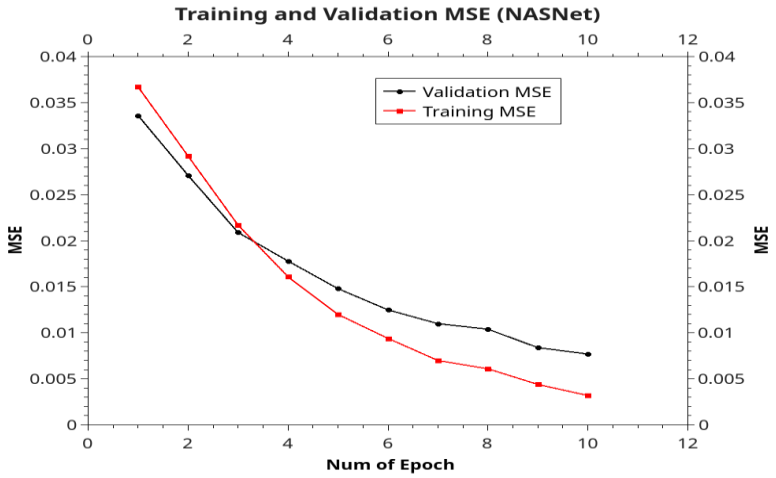

Figure 6. Training and validation MSE (Mean Squared Error)

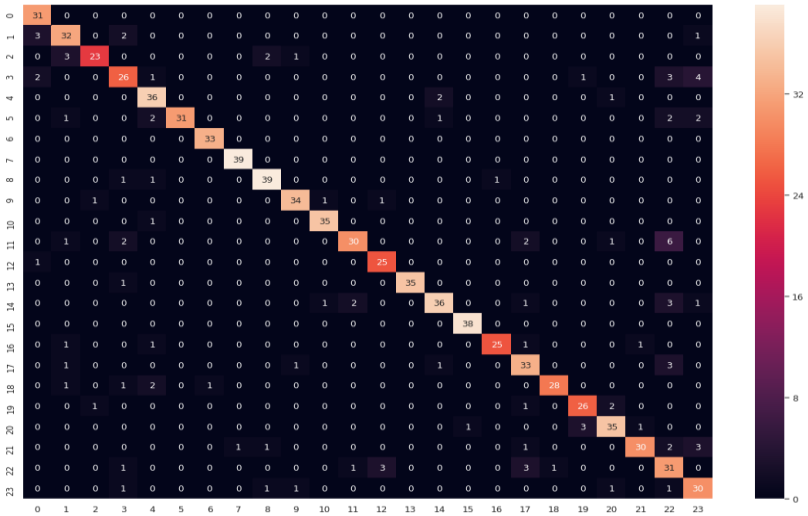

Figure 7. Confusion matrix

	precision	recall	f1-score	support
0	0.84	1.00	0.91	31
1	0.80	0.84	0.82	38
2	0.92	0.79	0.85	29
3	0.74	0.70	0.72	37
4	0.82	0.92	0.87	39
5	1.00	0.79	0.89	39
6	0.97	1.00	0.99	33
7	0.97	1.00	0.99	39
8	0.91	0.93	0.92	42
9	0.92	0.92	0.92	37
10	0.95	0.97	0.96	36
11	0.91	0.71	0.80	42
12	0.86	0.96	0.91	26
13	1.00	0.97	0.99	36
14	0.90	0.82	0.86	44
15	0.97	1.00	0.99	38
16	0.96	0.86	0.91	29
17	0.79	0.85	0.81	39
18	0.97	0.85	0.90	33
19	0.87	0.87	0.87	30
20	0.88	0.88	0.88	40
21	0.94	0.79	0.86	38
22	0.61	0.78	0.68	40
23	0.73	0.86	0.79	35
micro avg	0.87	0.87	0.87	870
macro avg	0.88	0.88	0.88	870
weighted avg	0.88	0.87	0.88	870

Figure 8. Precision, recall and F1-score metrics

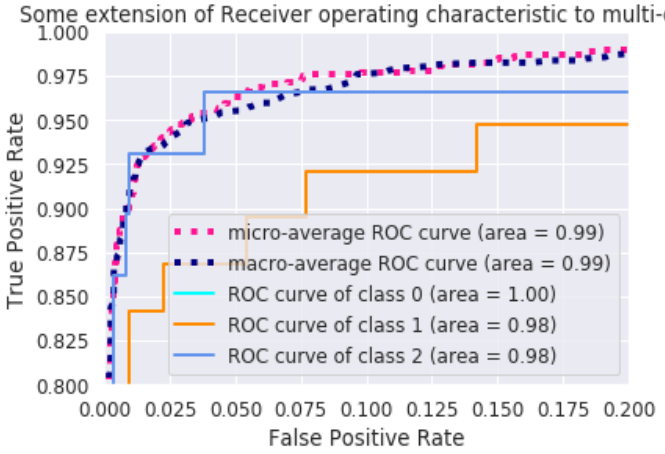

Figure 9. ROC curve and ROC area for each class

The ROC (Receiver Operator Characteristic) curve in Figure 9 graphically represents through curves (areas under the curves) the true positive and false positive rates in the model classification; the best categorization is achieved with the largest areas (blue).

5. CONCLUSION

In this paper, we show the development of a visual recognition system of the CSL in normal text in Spanish using a NASNET-type convolutional network. The CSL has static symbols that represent each of the letters used in the Spanish language, Colombia's mother tongue. The model was developed to be integrated into an assistive robot. The images used in the training (1000 in each of the 24 categories) were randomly mixed and resized to 256x256 pixels. For the network training, we used categorical cross-entropy loss, stochastic gradient descent, accuracy, and MSE. The design allows for the possibility of scaling to new symbols. The results showed a high performance, 88% accuracy measured on images unknown to the network (images not used in training and with unstructured environment). The model's confusion matrix shows a very low percentage of false positives, which is supported by the ROC curve of each category with average values of 0.99, indicating a very high accuracy. In addition, these values are consistent with the recall and f1-score metrics, which denotes an important contribution of the model, since it uses semi-structured images without previous processing, very close to real conditions, unlike previous studies where the images are conditioned in position, background and lighting, and use pre-processing.

ACKNOWLEDGEMENTS

This work was supported by the Universidad Distrital, the Universidad Surcolombiana, and the Universidad Nacional Abierta y a Distancia (UNAD). The views expressed in this paper are not necessarily endorsed by the above-mentioned universities. The authors thank the ARMOS research group for the evaluation carried out on prototypes.

REFERENCES

[1] N. Nagori and V. Malode, "Communication interface for deaf-mute people using microsoft kinect," *In International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT 2016)*, 2016. doi: 10.1109/ICACDOT.2016.7877664.

[2] A. R. Hasdak, et al., "Deaf-Vibe: A Vibrotactile Communication Device Based on Morse Code for Deaf-Mute Individuals," *2018 9th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pp. 39-44, 2018, doi: 10.1109/ICSGRC.2018.8657547.

[3] A. B. Jani, N. A. Kotak and A. K. Roy, "Sensor Based Hand Gesture Recognition System for English Alphabets Used in Sign Language of Deaf-Mute People," *2018 IEEE SENSORS*, pp. 1-4, 2018, doi: 10.1109/ICSENS.2018.8589574.

[4] Haq E. S., Suwardiyanto, D., M. Huda, "Indonesian sign language recognition application for two-way communication deaf-mute people," *In 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE 2018)*, pages 313–318, 2018. doi: 10.1109/ICITISEE.2018.8720982.

[5] Y. F. Admasu and K. Raimond, "Ethiopian sign language recognition using Artificial Neural Network," *2010 10th International Conference on Intelligent Systems Design and Applications*, pp. 995-1000, 2010, doi: 10.1109/ISDA.2010.5687057.
[6] L. Peluso, J. Larrinaga, and A. Lodi. Public policies on deaf education. comparative analysis between uruguay and brazil. Second International Handbook of Urban Education, vol. 2, no. 1, pp. 613–626, 2017, doi: https://doi.org/10.1007/978-3-319-40317-5_33.
[7] F. Lan, "The role of art education in the improvement of employment competitiveness of deaf college students," *In 2nd International Conference on Art Studies: Science, Experience, Education (ICASSEE 2018)*, pages 879–883, 2018, doi: https://doi.org/10.2991/icassee-18.2018.181.
[8] L. Vallejo, "El censo de 2018 y sus implicaciones en Colombia," *Apuntes del Censo*, vol. 38, no. 67, pp.9–12, 2019, doi: https://doi.org/10.19053/01203053.v36.n64.2017.6511.
[9] S. Goldin and D. Brentari, "Gesture, sign, and language: The coming of age of sign language and gesture studies," *Behavioral and Brain Sciences*, vol. 40, no. 46, pp. 1–60, 2017, doi: https://doi.org/10.1017/S0140525X15001247.
[10] Y. Motamedi, et al., "The cultural evolution of complex linguistic constructions in artificial sign languages," *In 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)*, pp. 2760–2765, 2017.
[11] J. Guerrero and W. Pérez, "FPGA-based translation system from colombian sign language to text," *DYNA*, vol. 82, no. 189, pp. 172–181, 2015, doi: http://dx.doi.org/10.15446/dyna.v82n189.43075.
[12] B. Villa, V. Valencia, and J. Berrio. Digital image processing applied on static sign language recognition system. Prospectiva, vol. 16, no. 2, pp. 41–48, 2018, doi: http://dx.doi.org/10.15665/rp.v16i2.1488.
[13] P. Kumar, et al., "Coupled hmm-based multi-sensor data fusion for sign language recognition," *Pattern Recognition Letters*, vol. 86, pp. 1–8, 2017, doi: https://doi.org/10.1016/j.patrec.2016.12.004.
[14] S. Hore, et al., "Indian sign language recognition using optimized neural networks," *Advances in Intelligent Systems and Computing*, vol. 455, no. 1, pp. 553–563, 2017, doi: https://doi.org/10.1007/978-3-319-38771-0_54.
[15] C. López, "Evaluación de desempeño de dos técnicas de optimización bio-inspiradas: Algoritmos genéticos y enjambre de partículas," *Tekhnê*, vol.11, no. 1, pp. 49–58, 2014.
[16] S. M. Kamal, et al., "Technical Approaches to Chinese Sign Language Processing: A Review," in *IEEE Access*, vol. 7, pp. 96926-96935, 2019, doi: 10.1109/ACCESS.2019.2929174.
[17] J. Gałka, et al., "Inertial Motion Sensing Glove for Sign Language Gesture Acquisition and Recognition," in *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6310-6316, Aug.15, 2016, doi: 10.1109/JSEN.2016.2583542.
[18] S. Chand, A. Singh, and R. Kumar, "A survey on manual and non-manual sign language recognition for isolated and continuous sign," *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 99–134, 2016, doi: 10.1504/IJAPR.2016.079048.
[19] F. Martínez, C. Penagos, and L. Pacheco, "Deep regression models for local interaction in multi-agent robot tasks," *Lecture Notes in Computer Science*, vol. 10942, pp. 66–73, 2018, doi: 10.1007/978-3-319-93818-9_7.
[20] F. Martínez, A. Rendón, and M. Arbulú, "A data-driven path planner for small autonomous robots using deep regression models," *Lecture Notes in Computer Science*, vol. 10943, pp. 596–603, 2018, doi: https://doi.org/10.1007/978-3-319-93803-5_56.
[21] E. B. Villagomez, et al., "Hand Gesture Recognition for Deaf-Mute using Fuzzy-Neural Network," *2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pp. 30-33, 2019, doi: 10.1109/ICCE-Asia46551.2019.8942220.
[22] J. Scott, S. Hansen, and A. Lederberg, "Fingerspelling and print: Understanding the word reading of deaf children," *American Annals of the Deaf*, vol. 164, no. 4, pp. 429–449, 2019, doi: 10.1353/aad.2019.0026.
[23] A. Lederberg, L. Branum, and M. Webb., "Modality and interrelations among language, reading, spoken phonological awareness, and fingerspelling," *Journal of deaf studies and deaf education*, vol. 24, no. 4, pp. 408-423, 2019, doi: 10.1093/deafed/enz011.
[24] Q. Zhang, et al,. "Myosign: enabling end-to-end sign language recognition with wearables," *In 24th International Conference on Intelligent User Interfaces*, pp 650–660, March 2019. doi: https://doi.org/10.1145/3301275.3302296.
[25] S. Fakhfakh and Y. Jemaa, "Gesture recognition system for isolated word sign language based on key-point trajectory matrix," *Computación y Sistemas*, vol. 22, no. 4, pp. 1415–1430, 2018, doi: 10.13053/CyS-22-4-3046.
[26] P. V. V. Kishore, et al., "Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pp. 346-351, 2016, doi: 10.1109/IACC.2016.71.
[27] K. Li, Z. Zhou, and C. Lee, "Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications," *ACM Transactions on Accessible Computing*, vol. 8, no. 2, pp. 1–23, 2016, doi: https://doi.org/10.1145/2850421.