ISSN 2007-9737

Siamese Convolutional Neural Network for ASL Alphabet Recognition

Atoany Nazareth Fierro Radilla, Karina Ruby Perez Daniel

Universidad Panamericana, Engineering Faculty,
Mexico

{afierro, kperezd}@up.edu.mx

Abstract. American sign language is an important communication way to convey information among the deaf community in North America and is primarily used by people who have hearing or speech impairments. The deaf community faces a struggle in schools and other institutions because they usually consist primarily of hearing people. Besides, deaf people often feel misunderstood by people who do not know sign language, for example, family members. In the last two decades, researchers have been proposing automatic sign language recognition systems to facilitate the learning of sign language, and nowadays, computer scientists have focused on using artificial intelligence in order to develop a system capable of reducing the communication gap between hearing and deaf people. In this paper, it is proposed a Siamese convolutional neural network for American sign language alphabet recognition. This siamese architecture allows the computer to reduce the high interclass similarity and high intraclass variations. The results show that the proposed method outperforms the state-of-the-art systems.

Keywords. Siamese network, CNN, ASL alphabet recognition, similarity learning, deep learning.

1 Introduction

ASL (American Sign Language) is an important communication way to convey information among deaf people. By visual signing, the brain processes linguistic information; this signing includes shape, movement, and placement of the hands, as well as facial expressions and body movements. ASL is not a universal language, each country has its own language, and in each region of each country, we can find dialects. Due to communication problems, it is very difficult for the deaf community the inclusion in school, job, and personal environments. Plenty of research works in automatic Sign Language Recognition (SLR) has been being published since two decades ago [1].

There are three types of automatic sign language recognition systems: 1) namely sentence; 2) words; 3) fingerspelling [1]. Fingerspelling (alphabetic sign language) is considered an essential part of learning sign language for new users and helps signers to perform signs for names of people, cities, and other words without known signs. There are some published works in which authors propose systems for ASL alphabet recognition [1, 10, 2, 5, 6, 7, 9, 8, 11, 12].

There are two important categories for ASL alphabet recognition, sensor-based and vision-based method. In the sensor-based approaches, the signer wears a special glove or sensor in order to present information of hand orientation, position and rotation, providing precise information. However, they are still too heavy and uncomfortable for daily use [1]. On the other hand, vision-based methods have been very popular because it does not need sensors attached to a human, and the low-cost cameras are commercially available. Vision-based methods use a digital image and apply image processing and machine learning techniques [1].

ASL alphabet recognition is a very difficult task due to high interclass similarities and high intraclass variations. In order to overcome this, in this paper, we propose to use a Siamese

Convolutional Neural Network (CNN) [3] in order to give the computer the ability of similarity learning and thus, reduce the interclass similarity and the intraclass variation of the non-linear representation of images of each sign of the ASL alphabet.

The rest of the paper is organized as follows: In Section 2 we present the related works; in Section 3 the proposed method is described; in Section 4 the experimental results are presented; in Section 5 we present the discussion about the practical application of the proposed scheme, in Section 6 we mention the future work and, finally in Section 7 we conclude this work.

2 Related Work

ASL alphabet recognition task is formulated as two subtasks: 1) feature extraction, and 2) multi-class classification. In [8], authors extracted features from color and depth images using Gabor Filters and then classify them using random forest, obtaining a 49% of precision. In [12], authors extracted shape, texture, and depth information from images and proposed a Superpixel Earth Mover's Distance (SP-EMD) to measure the distance between features of images.

Then, a template matching technique was utilized for sign classification, achieving a 75.8% recognition rate. Another related work was [6], where a Volumetric Spatiograms of Local Binary Pattern (VS-LBP) was used for extracting features and using a Support Vector Machine (SVM) an accuracy of 83.7% was achieved. In [7], features from depth images were extracted and classified them using random forest, getting an 81.1% of accuracy. In [5, 2], authors used depth images in order to recognize 24 classes of ASL alphabet using random forest, obtaining an accuracy of 87% and 90% respectively.

These approaches, as mentioned above, rely on two separated sub-tasks, feature extraction, and feature classification, where extracted features are well known as handcrafted features, due to the human intervention. The result of this separation produces a "decoupling phenomenon", where some important information for classification is missing in the feature extraction process. CNN networks have the advantage of doing both feature extraction and classification. Convolutional layers are responsible for obtaining non-linear representations of images (feature extraction), and Fully-Connected (FC) layers encode and classify these representations. In [10], a CNN was introduced, which has two inputs, one of them was for color images, and the other was for depth images.

Before fully connected layers, the representation of color and depth images are concatenated into one for classification, achieving 80.34% of accuracy. In [11], it is proposed a novel multi-view augmentation strategy, wherefrom only one depth image, and a 3D point cloud is obtained, then, additional cameras are set up and oriented to the point cloud with different perspectives. Finally, a set of additional views are generated from those distributed virtual cameras. In [1], authors proposed to use depth images captured by Microsoft Kinnect sensor and extract features from them using PCANet, and then these features are classified using Support Vector Machine (SVM), obtaining an 84.5% accuracy.

3 Proposed Method

One of the biggest challenging tasks in ASL alphabet recognition, as mentioned above, is the high interclass similarities and the high intraclass variance. In this paper, we propose a siamese architecture which can overcome these two problems performing a similarity learning and thus, reducing the interclass similarities and the intraclass variance among images.

For experiments, at first, we used small Siamese network architectures, for example, one architecture was composed of 4 convolutional layers and 1 fully connected layer, but this architecture was overfitted, and despite of having used a high Dropout rate, the network did not converge. We conclude from this experiment that the last feature maps were too small, and it was difficult for the network to have good learning.

Thus, we decided to increase the number of convolutional layers to 6 and to conserve the size of the feature maps using paddings, as well as to increase the number of dense layers due to they are responsible for encoding; this architecture

achieved a validation accuracy of 91%. This value of accuracy was too small, so we decided to add two more convolutional layers as well as to increase the number of neurons of the last dense layer. The proposed scheme was selected because it showed a better performance compared to the rest of the experimental architectures.

The proposed Siamese architecture is composed of two identical (siamese) convolutional neural networks sharing their parameters (weights and bias). Each of these two CNNs is compound by 8 convolutional and 3 fully-connected (dense) layers, as shown in Fig. 1.

A pair of images are presented as inputs, where this pair of images can be positive (images belonging to the same class) or negative (images belonging to different classes). These images are fed to convolutional layers that are responsible for feature extraction, such as color, texture, shape, edges, and orientations. Unlike CNN-based systems for image classification, dense layers of the proposed scheme carry out image feature encoding only, instead of encoding-classification. This encoding is fed to the contrastive loss where a similarity learning is performed. This similarity learning uses the distances between each pair of feature vectors generated by the last dense layer, obtaining as output a score that measures the similarity or dissimilarity between the pair of images (positive and negative, respectively). The detailed architecture of the proposed network is shown in Table 1.

Table 1. Detailed proposed CNN architecture

Layer (type)	Output shape	Param #
Convolution	64x64x16	448
Convolution	64x64x32	4,640
Max pooling	32x32x32	0
Convolution	32x32x32	9,248
Convolution	32x32x64	18,496
Max pooling	32x32x32	0
Convolution	16x16x64	39,928
Convolution	16x16x128	73,856
Max pooling	32x32x32	0
Convolution	8x8x128	147,584
Convolution	8x8x256	295,168
Batch Normalization	8x8x256	1024
Flatten	16,384	0
Dropout(0.5)	16,384	0
Dense	512	8,389,120
Dense	1024	525,312
Dense	4096	4,198,400

3.1 Similarity Learning

As we mentioned above, a pair of images (A and B) are fed into the networks; we proposed to use 64x64x3 images to reduce the computational cost. Each network generates a 4096-Dimensional feature vector (f(A) and f(B), respectively). Every CNN architecture for image classification is compound by convolutional layers for feature extraction and dense layers for encoding and classification, where the number of neurons in the last dense layer is equal to the number of classes.

In this case, the last dense layer of the proposed architecture consists of 4096 neurons because it is necessary to have a high-dimensional image representation to reduce the interclass similarities. In order to perform a similarity learning, first, the distance between the encoding of image A (f(A)) and image B (f(B)) is obtained as follows:

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(f(A)_i - f(B)_i)^2}, \qquad (1)$$

where D(.) is the distance between f(A) and f(B). If equation 1 is small, it means that A and B belong to the same class and vice versa. The contrastive loss is responsible for similarity learning and is defined as:

$$L = \frac{1}{2}lD^2 + \frac{1}{2}(max(0, m - D))^2, \qquad (2)$$

where l is a binary label indicating if A and B belong to the same class ($l = 1$) or not ($l = 0$); m is a margin selected for dissimilarity images (m must be greater than zero).

As can be observed from equation 2, the distance between two images of the same class

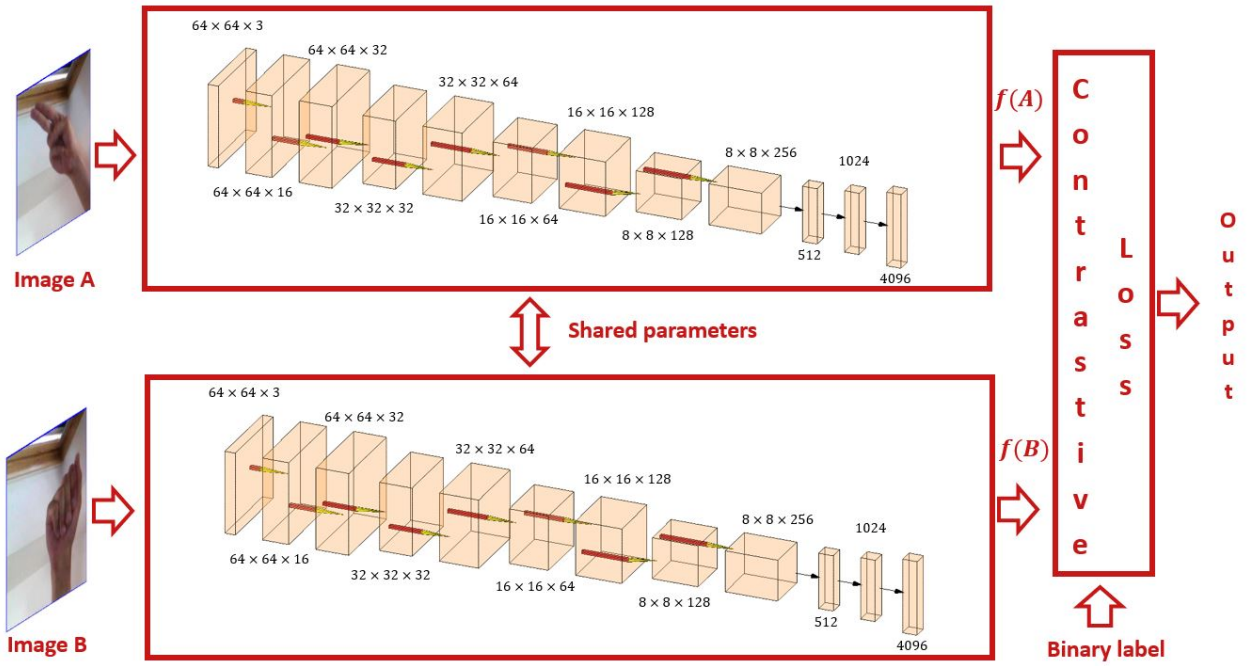

Fig. 1. The proposed architecture consists of two identical CNN which are sharing their parameters. Each network gets a representation of the input image and then they are fed into the contrastive loss for similarity learning. The output of the Siamese architecture is a score that indicates the similarity of the image pair

must be small, and for images belonging to different classes, the distance must be large. Thus, the networks generate codes for every image so that those who belong to the same class will have a small distance and vice versa. As a result, the large interclass similarity and the large intraclass variations are reduced, improving the classification rate of the ASL alphabet.

4 Experimental Results

The dataset we used for this paper is a sub-set from ASL Alphabet [4] dataset from Kaggle. This dataset consists of 26 ASL alphabet signs (from A to Z) and 3 classes labeled as “SPACE”, “DEL” and “NOTHING”, which according to the authors of the dataset, these are very helpful for real-time applications.

Something that is important to mention is that in this dataset, “J” and “Z” are considered static signs.

The subset used in this paper is compound by 8,700 random images (10% of the whole dataset). Before training, using this number of images, it was generated a set of 14,732 pairs of images (7,366 positive pairs and 7,366 negative pairs) from which 1,102 was used for testing (551 positives and 551 negatives). As we can see, using only 8,700 images, the number of training samples increased to 14,732.

The training was done using Keras and Tensorflow as frameworks on the Google Colab platform with a single 16GB Nvidia Tesla P100 GPU. After 30 epochs, the training loss and training accuracy were 0.0164 and 0.9870, respectively, and achieved a validation loss and a validation accuracy of 0.0245 and 0.9764, respectively. In Fig. 2, we can observe some classification results of the proposed scheme.

In Fig. 3, we present the training and validation curves, where we can observe there is any indication of overfitting due to we have implemented a Dropout of 50% in the flatten stage of the network. The effect of Dropout is like we were using different networks at each

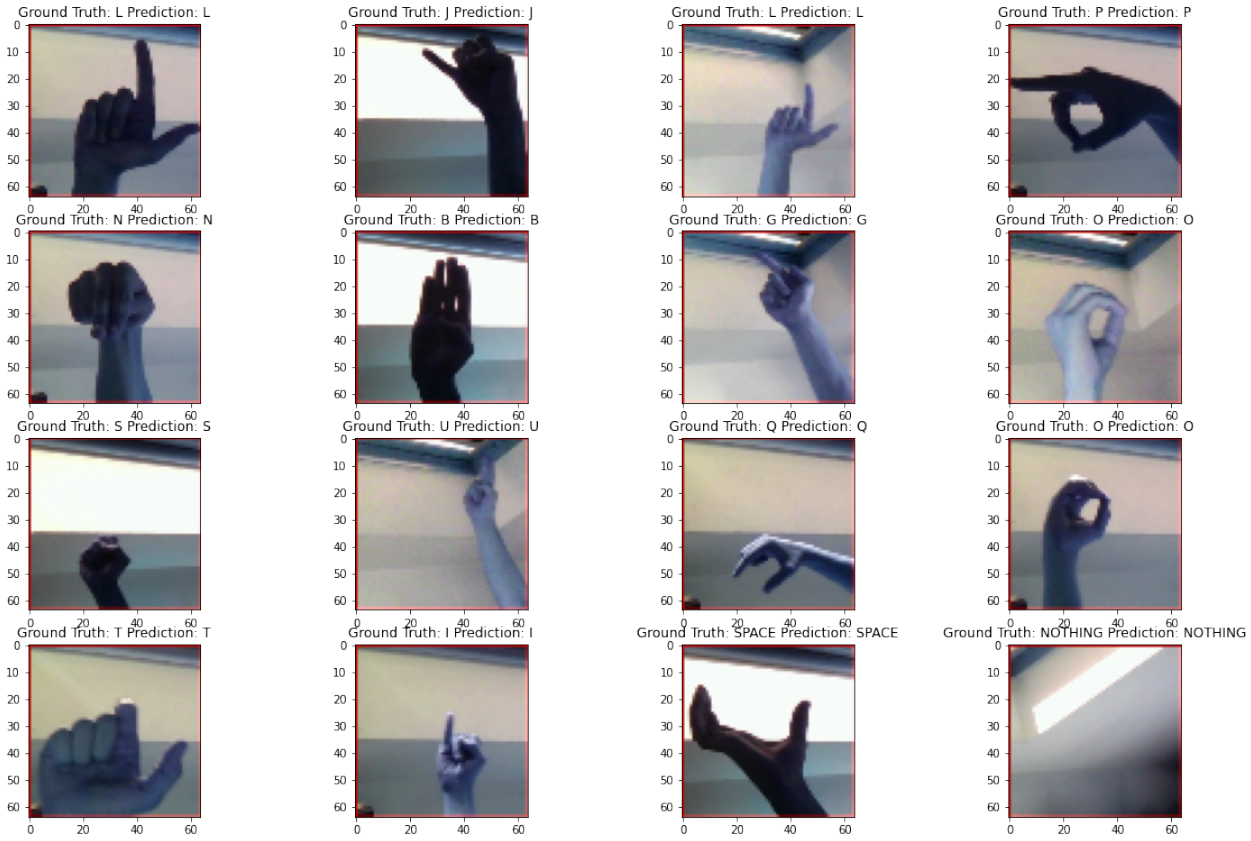

Fig. 2. ASL alphabet classification results using the proposed scheme

epoch because we randomly drop neurons at a rate of 50% in our case. This fact allows us to generalize the learning, getting a similar result in the validation set.

In order to evaluate the classification performance, we compute the confusion matrix shown in Fig. 4. The confusion matrix is a performance measurement for classification problems. It can be seen from Fig. 4 that the proposed scheme is doing an excellent performance on classifying the 29 classes.

We have used the accuracy, precision, and recall metrics to provide an evaluation in a quantitative manner. The results of these metrics are shown in Fig. 5. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations; recall, on the other hand, is the ratio of correctly predicted positive observations to all observations in the actual class.

From Fig. 5, we can observe that for the sign “M” and “N”, the proposed scheme achieved 93% and 85% of accuracy, respectively, and for the pair “R” and “U” achieved 86% and 85%, respectively. These values of accuracy were lower compared to the rest of the alphabet. This is because the sign for these letters is very similar (as shown in Fig. 6), and despite of having used a Siamese architecture, it remains some level of interclass similarity.

However, the average classification performance of the proposed method achieved an accuracy of 95%.

The proposed scheme was compared to published works where authors propose some other techniques for the same purpose but using different types of images (RGB and depth images)

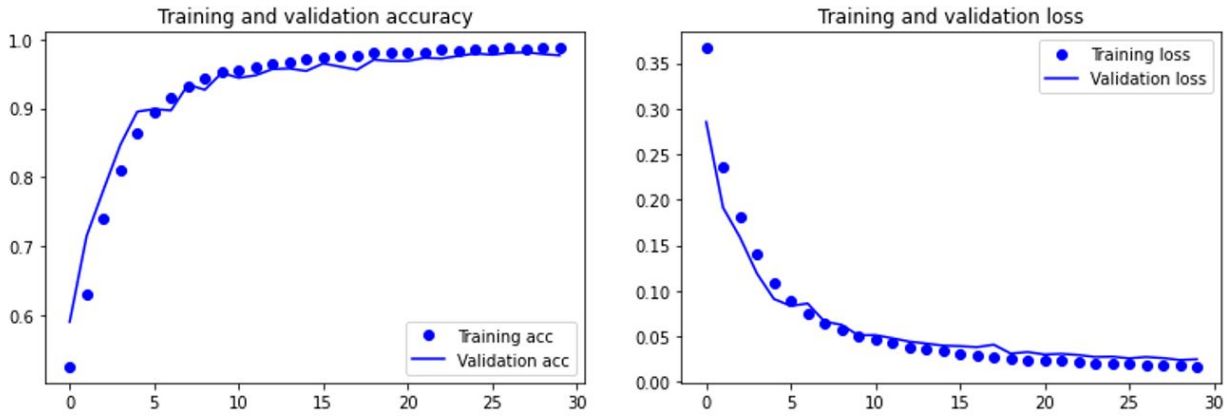

Fig. 3. Training curves. We can observe from these curves that the network is not overfitted

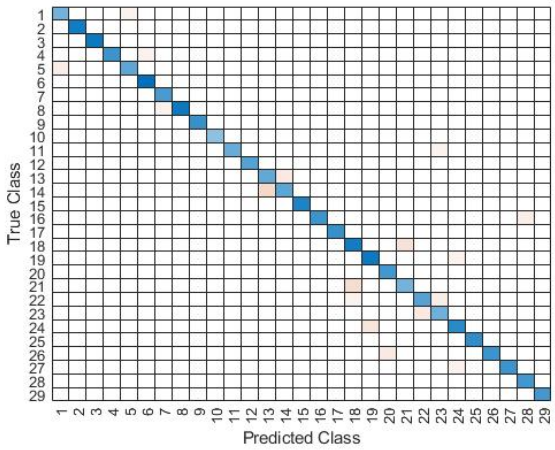

Fig. 4. Confusion matrix of the classification results using the proposed Siamese scheme

and different datasets. The results of this comparison are presented in Table 2.

5 Discussion

In this paper, we have proposed a system for ASL alphabet recognition which can help either hearing or no hearing people to learn sign language. The ASL language combines, as we mentioned above, hand movements and facial expressions.

Table 2. Comparison of the proposed method to published works

Method	Accuracy [%]
Aly et al. [1]	84.5
Ameen and Vadera. [10]	80.3
Dong et al. [2]	90
Kuznetsova et al. [5]	87
Maqueda et al. [6]	83.7
Nai et al. [7]	81.1
Pugeault and Bowden [8]	49
Tao et al. [11]	84.7
Wang et al. [12]	75.8
Proposed	96

In order to perform a communication translator, it is necessary to use videos instead of images for word and sentence recognition instead of symbol classification.

6 Conclusion

Sign language is not only important for people who are deaf, but also for people who want to communicate with them. Nowadays, the deaf community faces struggle due to the

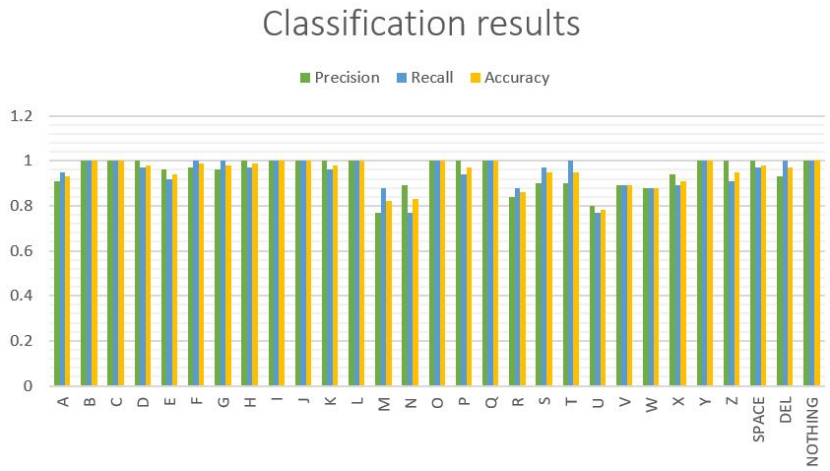

Fig. 5. Classification results per class of the ASL Alphabet [4] dataset

Fig. 6. It still remains some level of interclass similarity between the encoding of pairs “M&N” and “R&U”

communication gap that exists between hearing people and them. It is very important to develop a system for sign language translation to overthrow this communication wall.

In this paper, we propose a system to carry out the simplest task in ASL recognition, which is ASL alphabet recognition. One of the most challenging tasks in this field is the high interclass similarity and high intraclass variation in ASL alphabet recognition. Then, our hypothesis was to obtain image encoding where those belonging to the same class should be separated by a small distance (low variation) and at the same time by a large distance (low similarity) from those who belong to a different class. Therefore, we propose a Siamese architecture which uses two identical CNN. Experimental results show that our hypothesis is correct since we achieved to reduce the interclass similarity and intraclass variation, with some poor results in two pairs of classes. However, in general, we considered the proposed scheme performed well at classifying. The comparison presented in this paper shows that our neural architecture outperforms the published work in the literature.

7 Future Work

The results show that the proposed scheme outperforms the published work, despite we obtain not so good results in two pairs of images. Our future work will be to try to reduce the interclass similarity between the pair “M” and “N” and “R” and “U”. As well, we are planning to move one step forward and develop a system for a real-time ASL recognition system, including the movement for “J” and “Z”. In this case, we need to work with videos instead of images. In addition, we expect to develop these systems in a mobile device.

Acknowledgments

We would like to thank Universidad Panamericana for all the given support to accomplish this research.

References

1. **Aly, W., Aly, S., & Almotairi, S. (2019).** User-independent american sign language alphabet recognition based on depth image and pcanet features. *IEEE Access*, Vol. 7, pp. 123138–123150.
2. **Cao Dong, Leu, M. C., & Yin, Z. (2015).** American sign language alphabet recognition using microsoft kinect. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 44–52.
3. **Fierro, A., Nakano, M., Yanai, K., & Perez, H. (2019).** Siamese and triplet convolutional neural networks for the retrieval of images with similar contents. *Informacion Tecnologica*, Vol. 30, No. 6, pp. 243–254.
4. **Kaggle (2020).** Kaggle homepage. [Online available]: `https://www.kaggle.com/grassknoted/asl-alphabet`. [Accessed: 20/06/2020].
5. **Kuznetsova, A., Leal-Taixé, L., & Rosenhahn, B. (2013).** Real-time sign language recognition using a consumer depth camera. *2013 IEEE International Conference on Computer Vision Workshops*, pp. 83–90.
6. **Maqueda, A. I., del Blanco, C. R., Jaureguizar, F., & García, N. (2015).** Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns. *Computer Vision and Image Understanding*, Vol. 141, pp. 126–137.
7. **Nai, W., Liu, Y., Rempel, D., & Wang, Y. (2017).** Fast hand posture classification using depth features extracted from random line segments. *Pattern Recognition*, Vol. 65, pp. 1–10.
8. **Pugeault, N. & Bowden, R. (2011).** Spelling it out: Real-time asl fingerspelling recognition. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1114–1119.
9. **Sahoo, J. P., Ari, S., & Ghosh, D. K. (2018).** Hand gesture recognition using dwt and f-ratio based feature descriptor. *IET Image Processing*, Vol. 12, No. 10, pp. 1780–1787.
10. **Salem, A. & Vadera, S. (2017).** A convolutional neural network to classify american sign language fingerspelling from depth and colour images. *Expert Systems*, Vol. 34, No. 3, pp. 1–18.
11. **Tao, W., Leu, M. C., & Yin, Z. (2018).** American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, Vol. 76, pp. 202–213.
12. **Wang, C., Liu, Z., & Chan, S. (2015).** Superpixel-based hand gesture recognition with kinect depth camera. *IEEE Transactions on Multimedia*, Vol. 17, No. 1, pp. 29–39.

Article received on 18/06/2020; accepted on 20/07/2020.
Corresponding authors are Atoany Nazareth Fierro Radilla.