Comparison between handwritten word and speech record in real-time using CNN architectures

Javier O. Pinzón-Arenas, Robinson Jiménez-Moreno
Faculty of Engineering, Universidad Militar Nueva Granada, Colombia

Article Info

Article history:

Received Oct 10, 2019
Revised Feb 5, 2020
Accepted Feb 25, 2020

Keywords:

CNN
Faster R-CNN
Handwriting recognition
Speech recognition

ABSTRACT

This paper presents the development of a system of comparison between words spoken and written by means of deep learning techniques. There are used 10 words acquired by means of an audio function and, these same words, are written by hand and acquired by a webcam, in such a way as to verify if the two data match and show whether or not it is the required word. For this, 2 different CNN architectures were used for each function, where for voice recognition, a suitable CNN was used to identify complete words by means of their features obtained with mel frequency cepstral coefficients, while for handwriting, a faster R-CNN was used, so that it both locates and identifies the captured word. To implement the system, an easy-to-use graphical interface was developed, which unites the two neural networks for its operation. With this, tests were performed in real-time, obtaining a general accuracy of 95.24%, allowing showing the good performance of the implemented system, adding the response speed factor, being less than 200 ms in making the comparison.

Corresponding Author:

Robinson Jiménez-Moreno,
Mechatronics Engineering Program, Faculty of Engineering,
Universidad Militar Nueva Granada,
Carrera 11 # 101-80, Bogotá D. C., Colombia.
Email: robinson.jimenez@unimilitar.edu.co

1. INTRODUCTION

The implementation of pattern recognition applications in different types of data, such as signals from sensors [1], audio [2, 3], or even in images [4, 5], has been growing exponentially, giving way to the creation of a great variety of techniques to cover each type of pattern. Within these techniques, there is an area called deep learning [6], which contains robust pattern recognition methods such as recurrent neural networks [7, 8]. That are mainly used for speech recognition and writing, as well as deep belief neural networks [9], used for image recognition and natural language.

Another method of deep learning that has been evolving, mainly since early 2012 [6], is the convolutional neuronal network (CNN) [10], which was originally used to recognize only patterns in images. However, due to the demonstration of its capability to recognize up to 1000 different categories [11] and to support a large number of hidden layers, which even improved its performance [12, 13], its application has been extended to different fields, being used for speech recognition [14, 15], electromyographic signals [16, 17], and even as a means of human-machine interaction [18, 19]. With respect to speech recognition, CNNs have been applied successfully in different developments, such as the one shown in [20], where deep CNN architectures are compared against a deep neural network for large-scale speech tasks, having a relative improvement of 12% to 14%. In another development, this technique is used to recognize 6 different languages, obtaining a word error rate of 11.8% [21]. These works usually use phonemes to recognize the words, however, for an application where not the whole of a language is used, but some of the words, extracting the phonemes would result in a very complex task for the simplicity of the work, for this reason, as part of the contribution, this paper proposes using a CNN based on complete-word recognition, so that the input to the network is a whole processed audio, making its implementation much easier for the recognition of the words required.

On the other hand, regarding the location of objects in images, because CNN recognizes a single object in a total image, a variant was developed to allow it to identify several objects within the same image, generating a new variation of the network called region-based CNN (R-CNN) [22] that, in combination with a technique of identifying regions of interest (RoI), the network is able to locate and identify different objects but with a processing time of more than 10 seconds, making it very slow for any application in real-time. For this reason, other variations were implemented, reaching a network called Faster R-CNN [23] that, instead of using a separate algorithm to detect RoIs, has a region proposal network (RPN) with which it shares the weights and learning, which not only increases its robustness but also reduces the processing time the network takes to locate all the objects of interest in an image. However, although there are works that CNN have used to recognize handwritten words, such as those presented in [24, 25], the variation to locate them without the need for an additional algorithm has not been used, for that reason, this paper explores the use of a Faster R-CNN to locate 10 different handwritten words, to verify their operation.

This work, in addition to the aforementioned, seeks to implement a versatile application of truth check between what a user says and what they write, therefore, it is proposed to make the comparison of 10 different words, using a speech recognition system and one of location and identification of handwriting, united in a simple interface, which allows doing this task in real-time. This paper is divided into 4 sections, where section 2 describes the architectures implemented for speech recognition and handwritten word location and identification, and the interface developed for the task. In section 3, the results of the real-time test performed are presented. Finally, section 4 shows the conclusions reached.

2. METHODS AND MATERIALS

The development of this work is based on the recognition of 10 different words in the Spanish language that can be written by a user on a paper, which are: “Abajo” (Down), “Abra” (Open), “Arriba” (Up), “Avance” (Advance), “Cierre” (Close), “Dere” (shortening modification of Right), “Izqui” (shortening modification of Left), “No”, “Pare” (Stop), “Si” (Yes), additionally, the recognition of these is subject to what the user says by means of his voice. Taking into account this, the development depends on the implementation of two main functions, which are: one responsible for recognizing what the person has written and another responsible for processing the input audio, in such a way that it identifies the word that the user says. For this, two different CNN architectures are built and trained in order to execute said functions, and these are integrated into a graphic interface which is what the user uses. Its development is explained below.

2.1. CNN for handwriting recognition

The function oriented to recognize what the user wrote has two parts that are the location of the text and its subsequent classification. The objective was to implement a single neural network that is capable of carrying out both parts, for this reason, it was proposed to implement a neural network based on regions, type Faster R-CNN that, thanks to its high response capacity, allows the development of algorithms capable of locating and recognizing words in real-time. Based on this, the architecture shown in Table 1 is proposed, which consists of small square filters of 5x5, 3x3, and 2x2, since in terms of resolution, the words do not have large size, giving the possibility that the network is able to learn both generalities and more specific details of each word. Likewise, zeros are added to the edges of the images that pass through the network by means of padding, so that if a trace is very close to the edge, the network is able to learn or take into account its characteristics. Additionally, coupled to the CNN, there is a second path, which is a region proposal network (RPN), in charge of learning the regions or location of the words. These two are joined by means of a region of interest pooling (RoI-Maxpooling), which allows to have dynamic sizes of the incoming feature maps to this layer, i.e. depending on the box located in the RPN, it performs the corresponding downsampling to obtain the required input size in the next layer. A database containing a total of 1200 grayscale images is built, of which 1000 are used to carry out the training of the network and 200 to validate it. Each image was taken directly by a webcam and contains the 10 words to be identified, which were manually labeled, as shown in Figure 1.

With this, the training of the neural network is performed and later its validation. In this last step, the network obtains 98.9% of average accuracy in terms of the identification of the words within the image, and a mean average precision (mAP) in the location of the bounding boxes of 99.23%. The average precision of each word is shown in Figure 2, with the word "Yes" being the least accurate, with 98%, mainly because the estimated box tended to be a little larger than the ground truth set.

Table 1. Faster R-CNN architecture

Input		20x30	

Convolution (ReLU+BN)		5x5	F=32
			S=1/P=2

Convolution (ReLU+BN)		5x5	F=32
			S=1/P=2

MaxPooling		2x2	S=2

Convolution (ReLU+BN)		2x2	F=96
			S=1/P=2

Convolution (ReLU+BN)		3x3	F=96
			S=1/P=1

MaxPooling		2x3	S=2

Convolution (ReLU+BN)		2x2	F=128
			S=1

Path 1	Path 2		

	RPN: RoI-Convolution (ReLU)	3x3	F=128
			S=1/P=2

	RPN: Fully-Convolution	1x1	F=8
			S=1

	Softmax RPN		
RoI-MaxPooling		-	S = 2
FC (ReLU+Dropout)		512	
FC (ReLU+Dropout)		1024	
FC		11	
Softmax Classification			

Note: S=Stride, P=Padding, F=Filters, BN=Batch Normalization, FC=Fully Connected

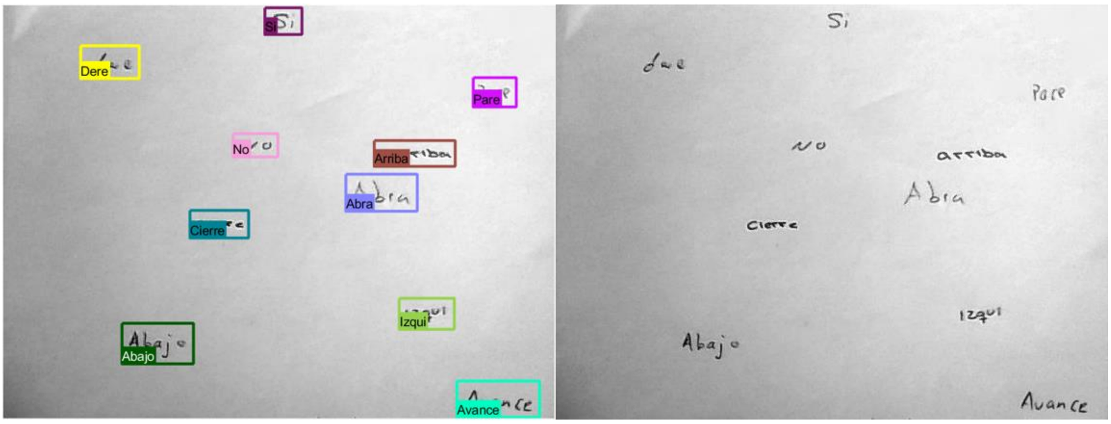

Figure 1. Sample of database with and without labels

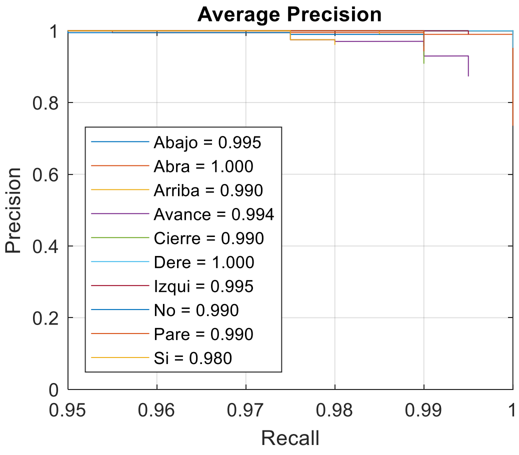

Figure 2. Behavior of the RoI detection at different recall levels in each category

2.2. CNN for handwriting recognition

To implement the speech recognition function of the user, first, the database for the training is prepared, in such a way that the dimensions of the network input are obtained. For this, it is obtained 115 audios, of length from 2 seconds to 16 kHz, of each word of different people in varied environments, i.e. controlled and not controlled or with noise, making a total of 1150 data in the database. However, as it is wanted to make audio acquisition continuously, there is a need to add an additional category, which in this case is called “Otros”, which contains different sounds of environment (people speaking, blows, etc.) and of the same users (strong breathing, coughing, etc.) to avoid that the network confuses these sounds with the words. From this category, 240 audios are recorded.

In order for the audios to be used in a CNN focused on speech recognition, processing of each audio is performed by means of the Mel Frequency Cepstral Coefficients (MFCC), in such a way that a feature map is obtained to be entered into the CNN. For the processing, a window of 20 ms is used with a shift of 10 ms, obtaining 12 coefficients per frame and a total of 199 frames. To have a clearer feature map, floor filter is applied to the map, and then obtain the first (Δ) and second (ΔΔ) derivative from the MFCC obtained, to acquire better characteristics of the sound and that the network has a greater possibility to learn the temporal variations of each word. An example of these maps is presented in Figure 3. Finally, the database of the feature maps obtained is divided, taking 100 maps per word and 200 from the category "Other" for training, and 15 per word and 40 from "Otros" for validation. Once the database is obtained, CNN is proposed for the application as shown in Table 2.

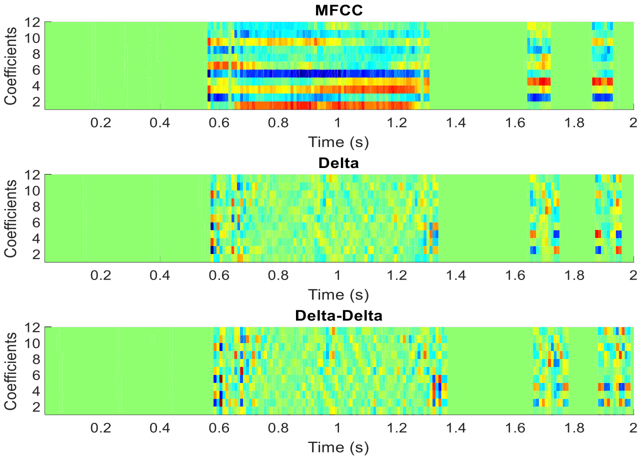

Figure 3. Feature maps of the word “Abajo

Table 2. CNN Architecture proposed

Input		12x199x3

Convolution
(+Batch Normalization)
	5x5	F=32
		S=1/P=2

Convolution	5x5	F=32
		S=1/P=2

MaxPooling	2x1	S=2x1

Convolution	3x3	F=64
		S=1/P=1

Convolution	3x3	F=64
		S=1/P=1

MaxPooling	2x3	S=2x2

Convolution	2x2	F=128
		S=1/P=1

Convolution	3x3	F=128
		S=1/P=1

MaxPooling	2x2	S=2x2
Fully Connected		512
Dropout		20% Disconnected
Fully Connected		2048
Dropout		20% Disconnected
Fully Connected		11
SoftMax		
Classification		

Since the recognition to be made is of the complete words, square filters are proposed, in such a way that the network learns the characteristics both in terms of the coefficients and the temporal variations. The network consists of 3 packets 2 convolutions plus a maxpooling, with the difference that in the first, a normalization layer is added in the first convolution and downsampling is made only in the coefficients, so that in the following convolutions, the temporal features are maintained. Likewise, padding is added to the volumes, in such a way that the characteristics found in the first and last coefficients are taken into account at the moment that the learning filters pass over them.

Finally, the training of the network is carried out, with which an accuracy of 100% and 90% validation is obtained, as shown in Table 3. In this one, it can be observed that the words that got the most erroneous classifications were “Arriba”, “Avance”, “Cierre”, and “Izqui”, mainly when there was a lot of ambient noise with people talking, which causes the characteristics map to change to some extent, causing it to be confused with other words.

Table 3. Classification of each word audio for validation

Word	Abajo	Abra	Arriba	Avance	Cierre	Dere	Izqui	No	Pare	Si	Otros	Overall acc.
True positives	15	14	12	12	12	15	12	15	14	14	36	90%

2.3. Graphic user interface

To concatenate the operation of the two trained CNNs, a basic graphic user interface (GUI) is made, so that it is easy to use for any user. This interface is composed of a window that allows visualizing in real-time the word written by means of a webcam. Additionally, there are 4 buttons, among which 2 are to activate and deactivate the webcam, and the other 2, to activate or deactivate the data collection by means of a microphone. Finally, there is a box that shows the word said by the user, which changes color depending on 2 cases: green, if the word mentioned by the user is the same as the one he recognized in the image, or red, if the word does not match the one recognized in the image. For the interface to start working, both the camera and the audio input must be activated. An example of the interface is shown in Figure 4.

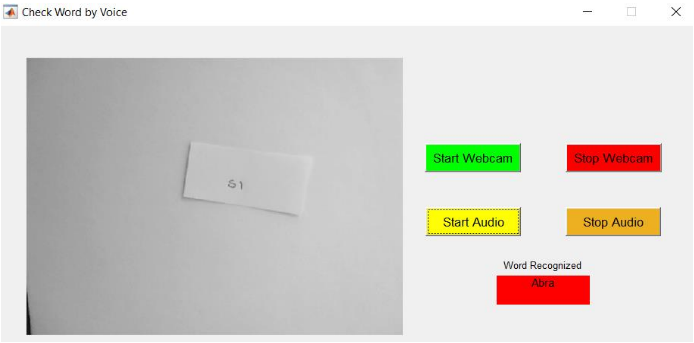

Figure 4. Graphic user interface

3. RESULTS AND ANALYSIS

Various real-time tests of the GUI are performed to see the performance of both the proposed network for speech recognition and the network specialized in finding the words. For this, a continuous recording algorithm is added, in such a way that once the audio is started, the data is taken continuously, however, in order to prevent environmental noise from being recognized, an amplitude threshold of the audio signal is used with respect to its average for a time of 200 ms. If the average of the amplitude exceeds the value of 0.1, it is determined that there is presence of voice, whereby the data is taken from a time ti=t-700 ms to tf=ti+2000 ms, where t is the current time, ti is the time of beginning of data collection and tf is the final time of data collection, obtaining an audio of 2 seconds. If the threshold is not surpassed, the audio data is not saved. With this, it is proceeded to perform the tests, where users write the words individually, placing them in any position of the image and even with a certain degree of inclination to increase the difficulty of their recognition. After it is located, the users say some of the words, either the one they wrote or another, in such a way that the interface shows them if what they say match with what they wrote. These tests are carried out with 6 users, each one with 7 tries, i.e. there were done 42 tests, obtaining the results shown in Table 4.

In each of the tests, the speech recognition network was able to correctly identify each word mentioned by the users. Regarding the identification of the words in the images, 95.24% of general accuracy was obtained, being able to recognize 20 words correctly and indicate an error of matching when the users named another word 20 times. Eight of these tests can be observed in Figure 5, where when the word was correctly classified, the identification confidence was greater than 90%, in other words, the network was "sure" that it was the word. On the other hand, the identification network of the words in the image had 2 erroneous classifications, that is, when the user mentioned a word and another was on the webcam, this was recognized as true, however, as shown in Figure 5.g and Figure 5.h, the confidence is even lower than 80%, which is solved with a confidence threshold, so that when a word is misclassified, if this threshold is not exceeded, it is taken as a negative.

Table 4. Results obtained from the tests

Right speech recognition	True positives	True negatives	False positives	False negatives	Overall accuracy
42	20	20	2	0	95.24%

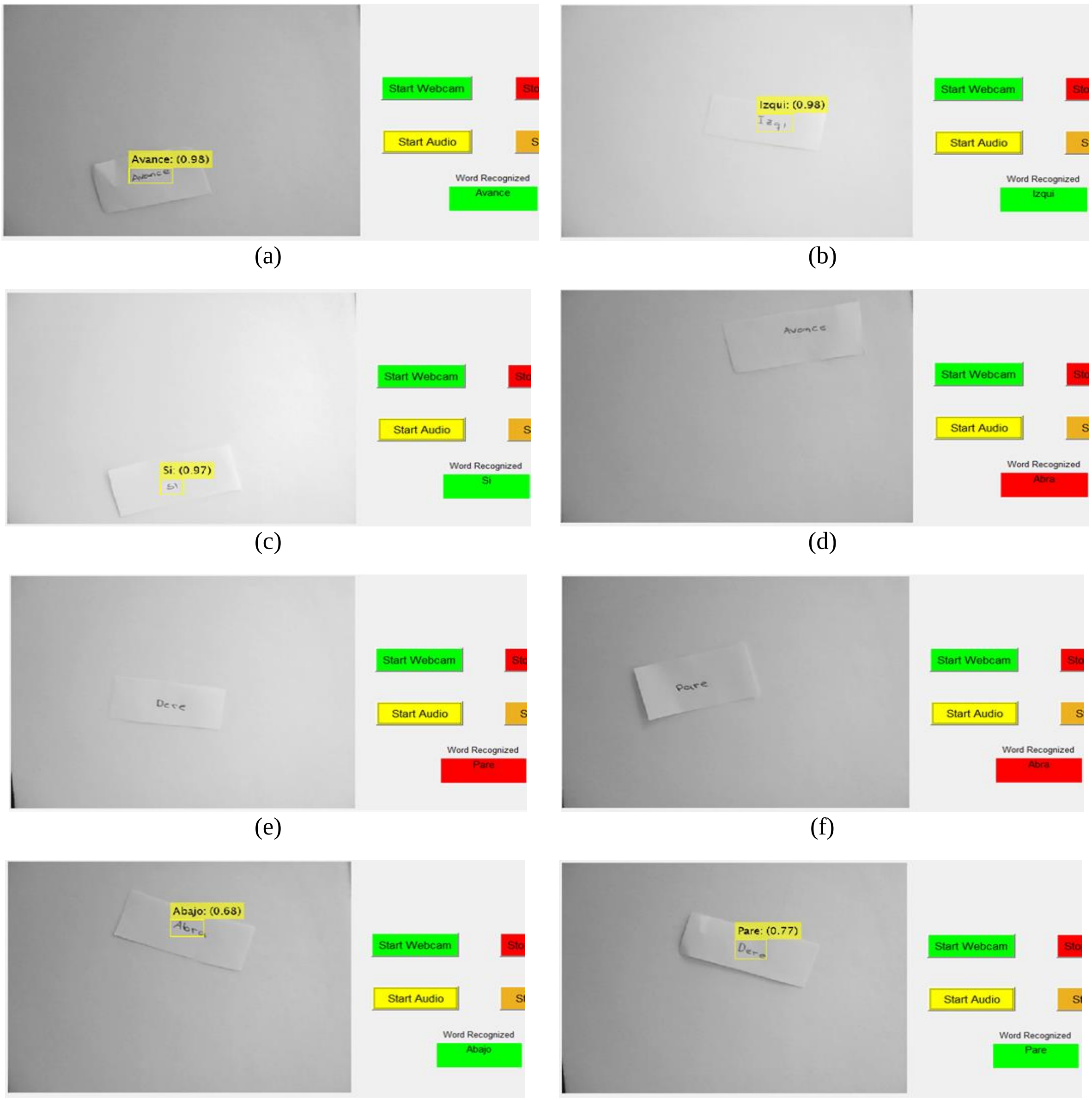

Figure 5. Samples of the test performed, (a, b, c) the written word is the one told by the user, (d, e, f) the written word differs from what the user said, (g, h) the written word is wrongly identified, passing it as correct compared with what was said by the user

In order to better understand how the used Faster R-CNN is behaving, the activations obtained by all the words in the last layer of the network are shown Figure 6. As can be seen, the network activates the most relevant regions of each word, even in the word “Cierre”, it focuses on the whole of it, where the red color means greater activation. On the other hand, in certain categories it activates the edges of the sheet where it has been written, mainly in the word “Si”, where possibly confuse the shadow with the section of the letter "i", however, not finding more features, the network does not identify it as a word. As for the processing time of the algorithm, once you have finished taking the audio, the process of obtaining the MFCCs, using the voice recognition network, using the Faster R-CNN and displaying the result in the graphic interface, it takes an average of 151 ms, i.e. once the user finishes saying the word, the algorithm identifies whether or not it matches with what is written almost instantly.

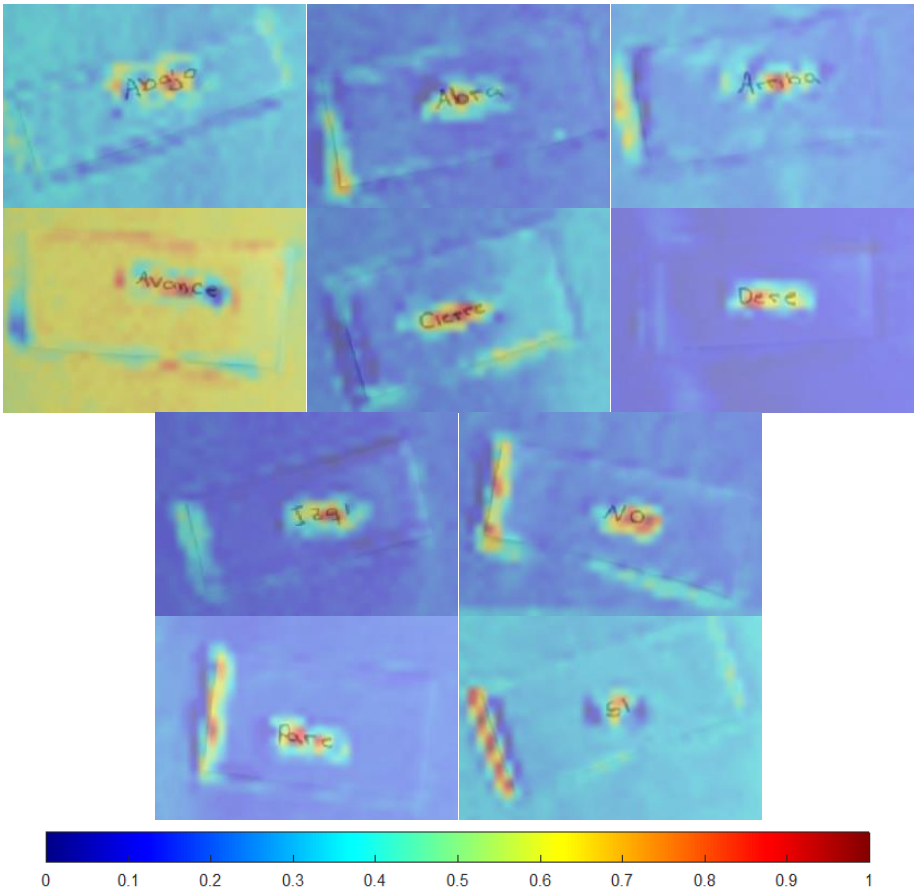

Figure 6. Faster R-CNN activations for some words

4. CONCLUSION

This paper described the implementation of a comparison system (executed in real-time) between handwritten and spoken words by a user so that each of the functions developed was implemented with different CNN architectures, which presented a validation accuracy of 98.9% and 90% for the recognition of handwriting and speech, respectively. On the other hand, when performing the tests in real-time, the speech recognition function did not have errors when recognizing each word that the user said, that is, it identified 100% of the 42 tests performed. Additionally, for the system of recognition and localization of the handwritten words, the difficulty of the tests was increased, not only by not controlling the ambient light, but also by placing them with a certain degree of inclination, in order to test its performance, obtaining an accuracy of 95.24% of the general system developed, where there were only 2 cases of false positives, i.e. when the user said any of the words, the system erroneously said that the written one matched, also, no false negatives were obtained during the tests performed. With this, it can be observed the very good performance of the general system implemented and the great capacity that the CNNs have to be used both in speech recognition using complete words and in location and recognition of handwriting, in addition to the fact that it is capable of making the comparison almost immediately regardless of whether 2 robust neural networks are being used, with a time of 151 ms. Taking into account the above, this system can be implemented and complemented for its application in different future works, for example, be used in countries where the traditional system of vote counting in electoral systems is still used, so that when the person says the number of votes for a candidate, the number written on the form must be the same, mainly to help avoid fraud in elections.

ACKNOWLEDGEMENTS

The authors are grateful to the Universidad Militar Nueva Granada, which, through its Vice-rectory for research, finances the present project with code IMP-ING-2290 (2017-2018) and titled "Prototipo de robot asistente para cirugía," from which the present work is derived.

REFERENCES

[1] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion recognition using bio-sensors: First steps towards an automatic system,” *Tutorial and Research Workshop on Affective Dialogue Systems,* pp. 36-48, 2004.

[2] G. D. Sree, P. Chandrasekhar, and B. Venkateshulu, “SVM based speech emotion recognition compared with GMM-UBM and NN,” *International Journal of Engineering Science and Computing,* vol. 6, no. 11, pp. 3293-3298, 2016.

[3] H. F. Pardede, A. R. Yuliani, and R. Sustika, “Convolutional neural network and feature transformation for distant speech recognition,” *International Journal of Electrical & Computer Engineering,* vol. 8, no. 6, pp. 5381-5388, 2018

[4] D. Cireşan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2012

[5] M. S. H. Al-Tamimi, “Combining convolutional neural networks and slantlet transform for an effective image retrieval scheme,” *International Journal of Electrical & Computer Engineering,* vol. 9, no. 5, pp. 4382-4395, 2019

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature,* vol. 521, no. 7553, pp. 436, 2015

[7] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” *2nd International Conference on Learning Representations (ICLR),* pp. 1-13, 2014.

[8] M. A. Jishan, K. R. Mahmud, and A. K. Al Azad, “Natural language description of images using hybrid recurrent neural network,” *International Journal of Electrical and Computer Engineering,* vol. 9, no 4, pp. 2932-2940, 2019

[9] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation,* vol. 18, no. 7, pp. 1527-1554, 2006

[10] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” *European Conference on Computer Vision,* pp. 818-833, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems,* pp. 1097-1105, 2012.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *International Conference on Learning Representations (ICLR),* pp. 1-14, 2015.

[13] C. Szegedy, V. Vanhouche, S. Loffe, J. Shlens, and Z. B. Wojna, “Rethinking the inception architecture for computer vision,” *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2818-2826, 2016.

[14] O. Abdel-Hamid, *et al*, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 22, no. 10, pp. 1533-1545, 2014

[15] S. Ganapathy and V. Peddinti, “3-D CNN models for far-field multi-channel speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 5499-5503, 2018.

[16] J. O. P. Arenas, R. J. Moreno, and R. D. H. Beleño, “EMG signal acquisition and processing application with CNN testing for MATLAB,” *International Review of Automatic Control,* vol. 11, no. 1, pp. 44-51, 2018.

[17] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, “Hand gesture recognition using compact CNN via surface electromyography signals,” *Sensors,* vol. 20, no. 3, pp. 1-14, 2020.

[18] C. Nuzzi, *et al.*, “Deep learning based machine vision: First steps towards a hand gesture recognition set up for collaborative robots,” *Workshop on Metrology for Industry 4.0 and IoT,* pp. 28-33, 2018.

[19] S. C. Hsu, Y. W. Wang, and C. L. Huang, “Human object identification for human-robot interaction by using fast R-CNN,” *2nd IEEE International Conference on Robotic Computing,* pp. 201-204, 2018.

[20] T. N. Sainath, *et al*, “Deep convolutional neural networks for large-scale speech tasks,” *Neural Networks,* vol. 64, pp. 39-48, 2015.

[21] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. Lecun, “Very deep multilingual convolutional neural networks for LVCSR,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4955-4959, 2016.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *IEEE conference on computer vision and pattern recognition,* pp. 580-587, 2014.

[23] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1137-1149, 2015.

[24] S. Sudholt and G. A. Fink, “PHOCNet: A deep convolutional neural network for word spotting in handwritten documents,” *15th International Conference on, Frontiers in Handwriting Recognition*, pp. 277-282, 2016.

[25] K. Zagoris, I. Pratikakis, and B. Gatos, “Unsupervised word spotting in historical handwritten document images using document-oriented local features,” *IEEE Transactions on Image Processing,* vol. 26, no. 8, pp. 4032-4041, 2017.

BIOGRAPHIES OF AUTHORS

Javier Orlando Pinzón Arenas was born in Socorro-Santander, Colombia, in 1990. He received his degree in Mechatronics Engineering (Cum Laude) in 2013, Specialization in Engineering Project Management in 2016, and M.Sc. in Mechatronics Engineering in 2019, at the Universidad Militar Nueva Granada-UMNG. He has experience in the areas of automation, electronic control, and machine learning. Currently, he is studying a Ph.D. in Applied Sciences and working as a Graduate Assistant at the UMNG with emphasis on Robotics and Machine Learning,
E-mail: u3900231@unimilitar.edu.co

Robinson Jiménez Moreno was born in Bogotá, Colombia, in 1978. He received the Engineer degree in Electronics at the Francisco José de Caldas District University-UD-in 2002. M.Sc. in Industrial Automation from the Universidad Nacional de Colombia-2012 and Ph.D. in Engineering at the Francisco José de Caldas District University-2018. He is currently working as a Professor in the Mechatronics Engineering Program at the Universidad Militar Nueva Granada-UMNG. He has experience in the areas of Instrumentation and Electronic Control, acting mainly in Robotics, control, pattern recognition, and image processing.
E-mail: robinson.jimenez@unimilitar.edu.co