VOL. 14, NO. 24, DECEMBER 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

COMPARISON BETWEEN VIOLA-JONES ALGORITHM AND SEMANTIC SEGMENTATION FOR FACE PARTS DETECTION

Javier O. Pinzón-Arenas and Robinson Jiménez-Moreno
Faculty of Engineering, Nueva Granada Military University, Bogotá D.C., Colombia
E-Mail: u3900231@unimilitar.edu.co

ABSTRACT

This paper presents the comparison between two face detection methods and their parts, which for this case are the two eyes and the mouth, which are the Viola-Jones (VJ) algorithm and the semantic segmentation based on convolutional neural networks (SegNet). To make the comparison, the training of the proposed SegNet is carried out using a database of previously labeled faces, to be later tested to verify its operation, where 97.55% of average accuracy and a mIoU of 76.64% are obtained. As for the VJ algorithm, an improved version for Matlab is used, which is able to detect the parts of the face even when it has an inclination of up to 20°. The tests are carried out with 10 images of the CelebA dataset, in such a way that each algorithm identifies the complete face, the right and left eye independently, and finally the mouth. In the event that any part of the face has been removed, the algorithm should not detect that section, since if it does it is counted as a false positive. In the tests, VJ obtained an overall accuracy of 79.38%, a low percentage compared to that obtained by SegNet, which was 97.97%. This allows seeing the capacity of the proposed network to identify the parts of the face and estimate when there is no information of any of the parts to be detected.

Keywords: viola-jones algorithm, semantic segmentation, SegNet, face parts detection.

1. INTRODUCTION

Nowadays, by means of machine vision and machine learning techniques, the possibility of creating applications that allow identifying and/or detecting objects in different scenes has been opened, even getting to recognize objects in images where the human being is not capable of doing it [1]. Within the multiple developments where these techniques are used, face detection is found. This detection is mainly done by means of the location of different characteristics of the face, being able to locate or discriminate within an image what is and what is not a face. One of the most widely used implementations for this purpose is the Viola-Jones algorithm [2], which through the use of cascade classifiers, allows the rapid detection of a face. Thanks to its performance, this architecture has been addressed to a large extent to make different modifications that increase even more the quality of detection [3], even managing not only to recognize the face, but also its most relevant parts [4], such as the eyes, mouth and nose. On the other hand, another variety of techniques based on Deep Learning [5] have been implemented in order to locate faces, such as those developed in [6] and [7], where different architectures of convolutional neural networks (CNN) are used to detect and locate the face in uncontrolled environments or even recognize some attributes of it.

Regarding the detection of parts or sections of the face, it is widely used in various developments, such as identification of people [8], facial verification [9], gender and age estimation [10], among others. For its implementation, the Viola-Jones algorithm can be applied, since in this case, an architecture of cascade classifiers is trained for each part, where the one made for the location of the face is the basis to be able to continue identifying the rest of the required sections [11], since if the face is not found, it would mean that the parts of the face do not exist either. However, with the great demonstrated performance of the CNN in the location of faces and the different variations that it has, it has begun to be used to detect the parts of the face, mainly by means of semantic segmentation [12]. An example of this is shown in [13], where a CNN cascade is used in such a way that the first stage is responsible for locating the landmarks of the face, and the second stage is responsible for carrying out the segmentation of the localized parts of the face. Since this type of structure Encoder-Decoder type allows a better characterization of the segmentation of objects, other developments have used this structure, as presented in [14], where they propose an architecture consisting of convolutions-deconvolutions, in order to identify and segment 7 categories between the background and parts of the face and hair, achieving average accuracies of approximately 85%.

This work focuses on the comparison between one of the most used methods for face detection and location of their respective parts, which is the Viola-Jones algorithm, and the one that has begun to take force in this field, which is semantic segmentation through CNN. To test them, 128x128 px images with faces with low resolution are used, between 50 px to 120 px high, with elimination of information, in order to check if they are able to know when a part of the face exists and when it does not. The parts to be detected are the face in general, each eye independently and the mouth. This development can be used to generate applications focused on the extraction of parts of the face for information collection, reconstruction of the same when its parts have been eliminated, among others.

The paper is divided into 4 sections, where section 2 describes the algorithms used to detect faces and their training if required. Section 3 shows the tests and results obtained. Finally, section 4 gives the conclusions reached.

2. Face Detection Algorithms

A. Neural Network for Semantic Segmentation

To perform the detection of the parts of the face by means of semantic segmentation, a convolutional neural network type Encoder-Decoder or SegNet [15] is proposed, with an architecture based on the VGG16 network [16], in such a way that it has a great depth in the network, allowing a greater number of features to learn, but entering images of lower resolution, which for this case are 128x128 px, to reduce the computational cost. Additionally, the advantages of SegNet are combined, which are the direct connection of each convolutions section between the encoder and the decoder, allowing to use each output image of the sections, i.e. to acquire the characteristics of both early layers of the network in the encoder and those estimated in the decoder and also, to reduce the translation invariance caused by each downsampling carried out, and, in relation to this, the improvement in the delineation of the edges.

The proposed architecture is shown in Figure-1, where it can be seen that each convolution has an additional set of layers, which are a batch normalization and a Rectified Linear Unit (ReLU), in such a way that it helps to reduce the internal shift caused by the convolution operation in the image, likewise, the interconnection of each MaxPooling with a MaxUnpooling, which, as mentioned above, allows to directly carry the output image of each encoder downsampling to the mirror section of the decoder, entering it to an upsampling to return the input volume to its initial size.

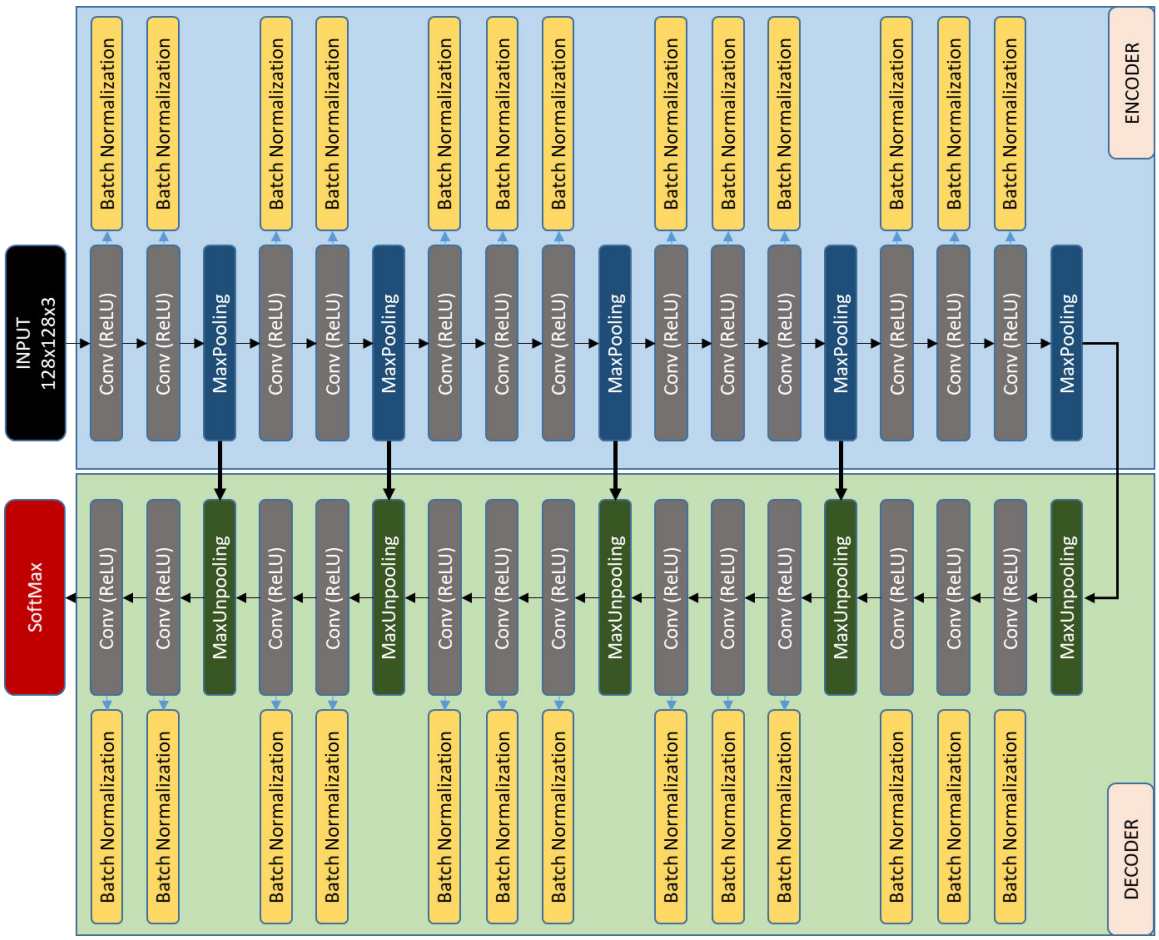


Figure-1. SegNet Architecture.

a) Dataset: To carry out the necessary training of the network, a database of our own is built, which contains RGB images of 128x128 px resolution, in JPG format. Due to the fact that in many cases faces are registered in low resolution in the images, it is decided not only to make the cut directly on the face, but to maintain a certain degree of background in the image in such a way that the size of the face varies between 50 px to 120 px within the image, which makes it more difficult to recognize certain parts of the face, since the number of pixels that make up are not many, for instance the eyes, causing some color blur with the rest of the face and shadows, and a loss of contour. In addition to this, the difficulty of recognizing the parts of the face and of the face itself is increased, by randomly eliminating the eyes (one or both) and the mouth, in this way, the algorithms need to be able to recognize, in the first place, that there is a face, and then, to find each of the parts. Some images were acquired from the CelebA dataset [7] to augment the database. Figure-2

shows some samples of the database constructed with its different characteristics.

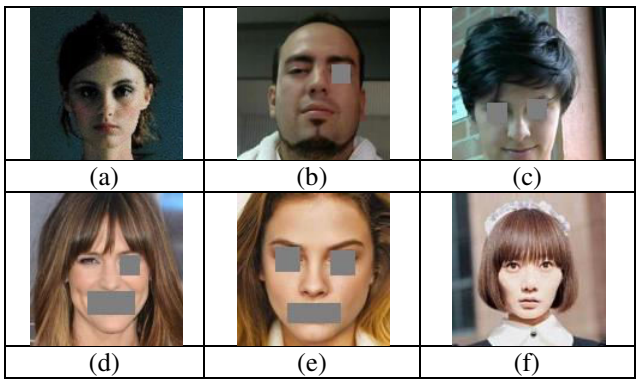


Figure-2. Dataset samples, where a) normal picture with shadows, b) face without an eye, c) without two eyes, d) without mouth and eye, e) without the two eyes and mouth and f) low resolution face (60 pixels height).

Once the database has been made, the corresponding labeling of the sections to be recognized is carried out. Since the application where it is wanted to use is focused on the location and removal of the parts of the face for further processing, the labeling is done not on the outline of the section, but leaving a certain shift that takes a small part of the face, as shown in Figure-3, added to the fact that in the low resolution faces, the eyes lose a lot of details.

Figure-3. Label Samples.

The parts or sections to be recognized are: the background, the face, the mouth and each eye independently. If in the face there are not any of the eyes or mouth, they are not labeled. In total, there are 225 images for the training, with variations of light, distance of the face and random removal of parts of the face. The database was built without a large amount of images to verify the effectiveness of the network to generalize without having a robust database.

b) Training and Validation: SegNet training is performed with the database implemented for 200 epochs, which results in the behavior shown in Figure-4, resulting in an accuracy of approximately 95%, however, when testing, the network is not able to correctly identify any part of the face, although it is well labeled, as shown in Figure-5.

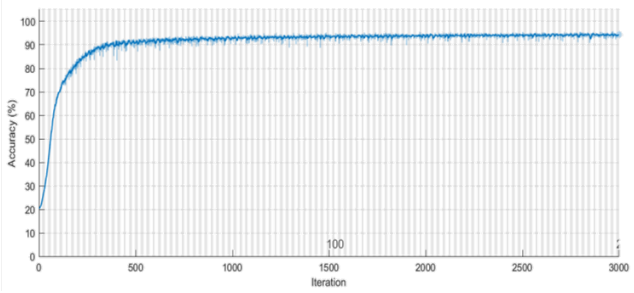


Figure-4. SegNet Training Behavior.

Figure-5. Original image, ground truth and SegNet output.

Although it learned quickly, its validation does not exceed an average of 39.38% between categories, having that, despite the fact that the face and background have more than 90% accuracy, the eyes and mouth have 0%, that is, they were not identified in any test image, being confused mainly with the face, as it is shown in the confusion matrix of Figure-6.

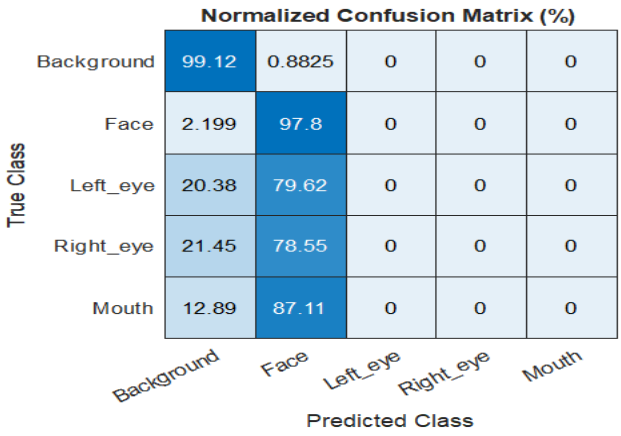


Figure-6. Confusion Matrix of the test.

The high percentage of inaccuracy of the network is due to the fact that the network is biased in learning the labels that are most found in the images. To know the frequency of appearance of the labels, the summation of all the pixels of each category of the complete database is done, obtaining the graph of Figure-7, where it is evident that the face and the background are the ones that appear most, while the parts of the face do not have a frequency greater than 0.03, making these labels relatively insignificant and the network biasing to those of greater value.

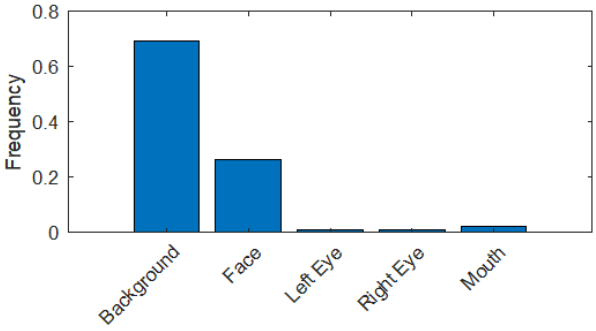


Figure-7. Category Frequency in the Database.

To solve this, a weight assignment is made in the classification stage of the network, in order to balance the recognition of each class. For this, the weights of each category are calculated by means of the inverse frequencies using (1).

$$w_i = \frac{1}{F_i} = \frac{1}{n_i/N} \quad (1)$$

Where *i* is the category, *w* the frequency weights, *n* the number of pixels of the current category, *N* the sum of the number of pixels of all the categories and *F* the frequency of presence of the category. With this, weights are assigned to each category, *w=1.4410* for the background, w=3.7812 for the face, *w=98.4207* for the left eye, *w=99.9119* for the right eye and *w=46.6857* for the mouth.

With the weights, the training of the network is again performed, with which the behavior shown in Figure-8 is obtained, where unlike the first training (Figure-4), its learning is more delayed, its learning is more delayed, but he manages to achieve the same training accuracy in the 200 epochs. With this training, the test is performed to verify that the network has a correct functioning, which is evidenced in Figure-9, where the network is able to label each category in a very approximate way to the ground truth, additionally, averaging the accuracies obtained in the confusion matrix shown in Figure-10, a mean accuracy of 97.55% is obtained, where, on this occasion, the parts of the face obtained more than 99% accuracy and a mean intersection over union (mIoU) of 76.64% of all the categories, in other words, the precision of the area obtained with respect to the expected area.

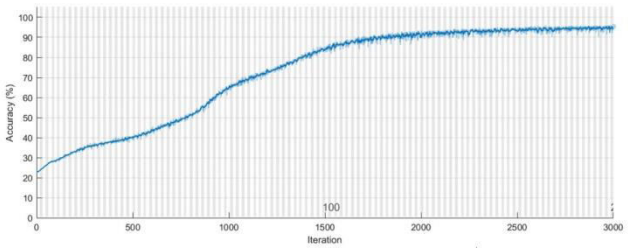


Figure-8. SegNet Training Behavior with Frequency Weighting.

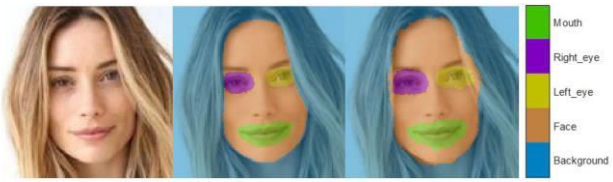


Figure-9. Original image, ground truth and SegNet output with frequency weighting.

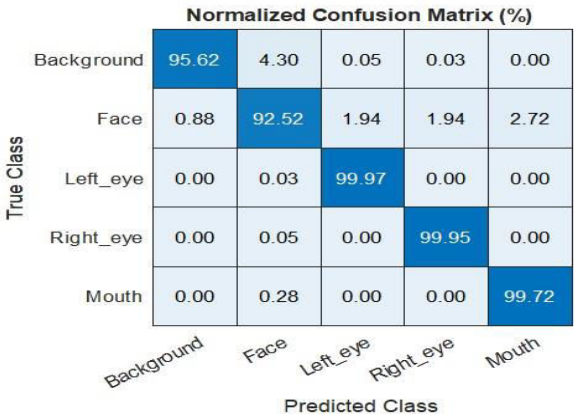


Figure-10. Confusion Matrix of the test with frequency weighting.

B. Viola-Jones Algorithm

To perform the comparison with the trained SegNet, an object detector by means of cascade classifiers based on the Viola-Jones algorithm [2] is used. For this case, use is made of the face detection toolbox implemented by Matlab® together with the improvements made by Masayuki Tanaka [17], which allow identifying faces with high degrees of inclination and with resolutions lower than those that the toolbox by default is capable of detecting. In this case, the algorithm is named "VJ".

3. RESULTS AND DISCUSSIONS

To make the comparison of the SegNet against the VJ for the location of the 3 parts of the face, 10 images of the CelebA dataset are chosen that have not been used for the training and have different levels of complexity, i.e. that there are faces front, with side movement, with degree of inclination with respect to the horizontal and that have different shades of light and shadows. These images are cut square and resized to a size of 128x128 px, which the parts of the face to identify were manually erased, so that each algorithm, apart from having to find the face, must identify whether or not each part of the face exists, that is, if the mouth is cut or erased, it should not indicate that there is a mouth on the face. In total, a test dataset of 80 images is created, with 10 images unchanged and 70 images modified (with erased face parts).

Figure-11 shows 2 examples of detection of the parts of the face, where in Figure-11a it can be seen that both the VJ and the semantic segmentation were able to correctly identify the location of the eyes and mouth, although SegNet cuts a part of the face at the top. In Figure-11b, the VJ is not able to identify the right eye,

because its visible section is small, while SegNet could detect it, although on the other hand, the network does not cover the entire mouth, possibly due to the use of dark-colored lipstick, since no training images with this lip color are found.

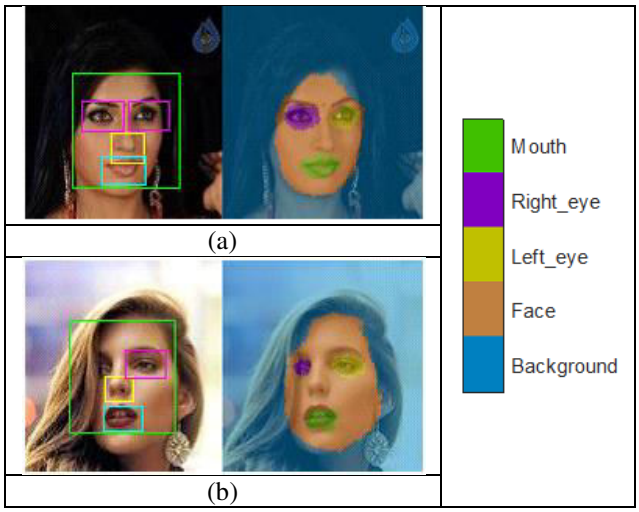


Figure-11. Comparison with test images between Viola-Jones (Left) and Semantic Segmentation (Right).

For the comparison, as the SegNet depends on the detected pixels, it is established that categories of eyes with less than 20 pixels detected, is assumed as negligible or unidentified, while for the mouth they are less than 40 pixels. This is mainly because when filtering, they are areas that are eliminated, additionally, the area obtained must be correctly located on the detected part, that is, if the area surrounds the entire section with a shift less than 60%, it will be graded with a weight of 1 (True Positive with right positioning), but if the area takes between 30% and 60% of the detected part, it will be qualified with 0.5 (True Positive with regular positioning), otherwise, it will be classified as 0 (False Negative) and if the area exceeds the pixel threshold when the section does not exist, it is rated as 0 (False Positive), otherwise, it will be 1 (True Negative). For the VJ, apart from locating the part of the face, it must locate the bounding box correctly, qualifying in the same way as the SegNet with respect to the section that takes the box.

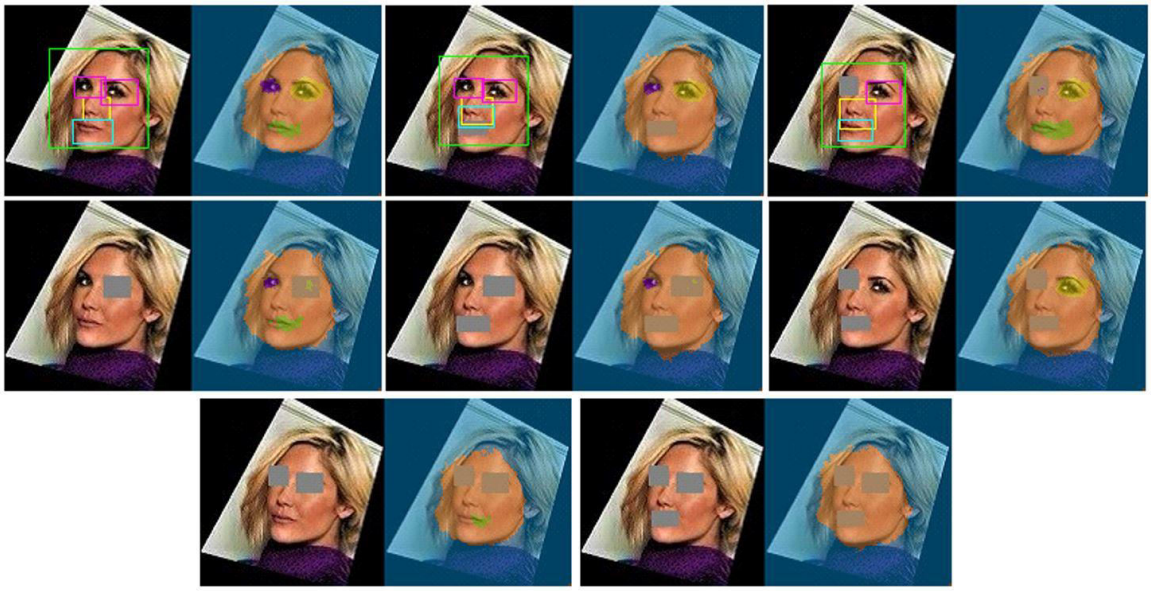
Figure-12 shows 2 examples of complete comparison, with all the modifications made to each image. Figure-12a shows how the two algorithms work with a face located front and without tilt, where, despite removing even the two eyes and the mouth at the same time, they are able to recognize the location of the face. However, the main problem with VJ is that it detects the mouth as it has been removed. On the other hand, the SegNet is able to know that there is no part of the face, except in two of the modified images. In the image where the left eye and the mouth are cut, the network identifies some pixels of the eye when it has been removed; however, the number of pixels detected is negligible. Similarly, it happens in the image that cut the 3 parts of the face, where some pixels of the mouth are located due to the beard, but the amount detected does not reach the threshold established to be qualified as a false positive.

Figure-12b shows the comparison of the algorithms with a more complex face, which, apart from having a degree of inclination, is not completely frontal. In this, the VJ is only able to locate the face in the first 3 images, and continues to detect the mouth when it has been removed, while in the other modifications, not being able to detect the face, it is not possible to know if there are parts of it. On the other hand, the SegNet was able to locate the face in all the images, even if it has taken a section of the hair, however, the mouth, although the network detects it, does not place it at all well, since it has a certain shift. Also, in some cases, it detects pixels of the left eye, although it has been removed, and in one of the cases, the number of pixels exceeds the detection threshold, causing it to become a false positive.

Table-1 shows the results obtained with the 10 images and their 8 variations, such as those shown in Figure-12. In general, it can be seen that the VJ algorithm failed mainly when the mouth was not in the image, as in the example of Figure-12a, which located the mouth in another section, even if it was not there, which generated a lot of false positives, making the true negatives only reach 27.5% in terms of the mouth and about 65% in general, reaching a consolidation with a overall accuracy of 79.38%.

Regarding the SegNet, where it most failed was when the mouth was found by removing the left eye, although sometimes the number of pixels found did not exceed the threshold, however, the network had approximately 98% overall accuracy detecting the parts of the face. This allows observing that the trained SegNet has a greater ability to detect face parts compared to the VJ, even when information on certain parts of the face is suppressed.

(a) Front face without inclination



(b) Face slightly moved to the side, with an inclination of 15°

Figure-12. Samples of modification of the images and comparison between the two algorithms.

Table-1. Comparison Test performed.

Modification	VJ				SegNet			
	Face	**L-Eye**	**R-Eye**	**Mouth**	**Face**	**L-Eye**	**R-Eye**	**Mouth**
Original	100%	90%	90%	100%	100%	95%	100%	100%
No Mouth	100%	90%	90%	20%	100%	100%	100%	100%
No R-Eye	100%	90%	90%	100%	100%	100%	100%	100%
No L-Eye	90%	90%	80%	90%	100%	90%	100%	85%
No L-Eye+Mouth	90%	90%	80%	50%	100%	100%	90%	100%
No R-Eye+Mouth	70%	70%	70%	20%	100%	100%	100%	100%
No L-Eye+ R-Eye	90%	90%	90%	90%	100%	100%	90%	90%
No L-Eye+ R-Eye +Mouth	70%	70%	70%	20%	95%	100%	100%	100%
True Positive Accuracy	88.75%	85.00%	85.00%	95.00%	99.38%	98.75%	97.50%	93.75%
	88.44%				97.75%			
True Negative Accuracy	-	85.00%	80.00%	27.50%	-	97.50%	97.50%	100%
	64.17%				98.33%			
Overall Accuracy	88.75%	85.00%	82.50%	61.25%	99.38%	98.13%	97.50%	96.88%
	79.38%				97.97%			

4. CONCLUSIONS

This work presented the comparison of two methods to locate the parts of the face, which were the algorithm of Viola-Jones (VJ) and a convolutional neuronal network type Encoder-Decoder (SegNet). The main motivation to make this comparison is presented in the need to be able to extract certain sections of the face, taking into account that there may or may not be information of some of the required sections. When performing the comparative analysis of the two algorithms using images with which the SegNet was not trained, it shows the great capacity of the neural network by detecting the parts of the face, even guessing at the non-location of the eliminated sections, with a high degree of accuracy of 97.97%, compared to the 79.38% obtained with the VJ.

Although the VJ algorithm is widely used for face detection, in images with low resolution it has complications when it comes to detecting the parts of the face, mainly because they cannot be easily distinguished because the pixels distort the sections. On the other hand, the proposed SegNet is able to identify the required section, in addition to the fact that it has a higher performance when the information of the parts of the face has been eliminated, even when the face is not completely in front and with a certain inclination, which for the VJ is presented as a critical point of detection failure, as shown in Figure-12b.

In addition to the comparison made, it could be observed that semantic segmentation does not necessarily require that the objects or parts to be recognized be labeled with borders, but that they are labeled taking areas from another part, as it was done in this work, where part of the face is taken to label the eyes and mouth, and thus detect not only the edges, but the area where they are located.

The capacity shown by the SegNet allows to open the field of application of this for future work, such as the identification of people in surveillance systems, where, although the person covers parts of his face to go unnoticed, the network is able to detect and locate his face, and extract the parts required for further analysis, especially because these systems tend to have low resolution image captures.

ACKNOWLEDGMENTS

The authors are grateful to the Nueva Granada Military University, which, through its Vice chancellor for research, finances the present project with code IMP-ING-2290 (2017-2018) and titled "Prototype of robot assistance for surgery", from which the present work is derived.

REFERENCES

[1] Nguyen, J. Yosinski and J. Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 427-436.

[2] P. Viola and M. J. Jones. 2004. Robust real-time face detection. International journal of computer vision. 57(2): 137-154.

[3] Y. Q. Wang. 2014. An Analysis of the Viola-Jones face detection algorithm. Image Processing On Line. 4: 128-148.

[4] M. Castrillón, O. Déniz, D. Hernández and J. Lorenzo. 2011. A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework. Machine Vision and Applications. 22(3): 481-494.

[5] Y. Le Cun, Y. Bengio and G. Hinton. 2015. Deep learning. Nature. 521(7553): 436.

[6] S. Yang, Y. Xiong, C. C. Loy and X. Tang. 2017. Face Detection through Scale-Friendly Deep Convolutional Networks. arXiv preprint arXiv: 1706.02863.

[7] Z. Liu, P. Luo, X. Wang and X. Tang. 2015. Deep learning face attributes in the wild. in Proceedings of the IEEE International Conference on Computer Vision. pp. 3730-3738.

[8] O. K. Manyam, N. Kumar, P. Belhumeur and D. Kriegman. 2011. Two faces are better than one: Face recognition in group photographs. in Biometrics (IJCB), 2011 International Joint Conference on, IEEE. pp. 1-8.

[9] N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar. 2009. Attribute and simile classifiers for face verification. in Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 365-372.

[10] E. Eidinger, R. Enbar and T. Hassner. 2014. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security. 9(12): 2170-2179.

[11] K. Vikram and S. Padmavathi. 2017. Facial parts detection using Viola Jones algorithm. in Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on, IEEE. pp. 1-4.

[12] R. Girshick, J. Donahue, T. Darrell and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580-587.

[13] A. S. Jackson, M. Valstar and G. Tzimiropoulos. 2016. A CNN cascade for landmark guided semantic part segmentation. in European Conference on Computer Vision, Springer, Cham. pp. 143-155.

[14] M. M. Kalayeh, B. Gong and M. Shah. 2017. Improving facial attribute prediction using semantic segmentation. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4227-4235.

[15] V. Badrinarayanan, A. Kendall and R. Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence. 39(12): 2481-2495.

[16] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.

[17] M. Tanaka. 2016. Face Parts Detection. [Online]. Available in: https://la.mathworks.com/matlabcentral/fileexchange/36855-face-parts-detection