

# Comparison of convolutional neural network models for user's facial recognition

Javier Orlando Pinzón-Arenas<sup>1</sup>, Robinson Jiménez-Moreno<sup>1</sup>, Javier Eduardo Martínez Baquero<sup>2</sup>

<sup>1</sup>Mechatronic Engineering, Faculty of Engineering, Universidad Militar Nueva Granada, Bogota, Colombia

<sup>2</sup>Engineering School, Faculty of Basic Sciences and Engineering, Universidad de los Llanos, Villavicencio, Colombia

## Article Info

### Article history:

Received May 10, 2023

Revised Jul 12, 2023

Accepted Jul 17, 2023

### Keywords:

Architecture comparison  
Biometric register  
Convolutional neural network  
Face recognition  
Transfer learning  
User identification

## ABSTRACT

This paper compares well-known convolutional neural networks (CNN) models for facial recognition. For this, it uses its database created from two registered users and an additional category of unknown persons. Eight different base models of convolutional architectures were compared by transfer of learning, and two additional proposed models called shallow CNN and shallow directed acyclic graph with CNN (DAG-CNN), which are architectures with little depth (six convolution layers). Within the tests with the database, the best results were obtained by the GoogLeNet and ResNet-101 models, managing to classify 100% of the images, even without confusing people outside the two users. However, in an additional real-time test, in which one of the users had his style changed, the models that showed the greatest robustness in this situation were the Inception and the ResNet-101, being able to maintain constant recognition. This demonstrated that the networks of greater depth manage to learn more detailed features of the users' faces, unlike those of shallower ones; their learning of features is more generalized. Declare the full term of an abbreviation/acronym when it is mentioned for the first time.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## 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

Face detection is one of the research topics that have remained in force in state of the art [1], gaining high relevance in applied systems such as access control to safe areas [2], attendance management systems [3], or young people recognition with some risk of vulnerability [4]. For face detection, the techniques used are very varied. For example, the use of binary descriptors [5], mechanisms based on three-way decisions [6], and alignment learning [7], all of them based mainly on image and video analysis [8]. One of the most efficient techniques for extracting features through images or videos focuses on deep learning algorithms.

Deep learning techniques have witnessed notable advancements in face identification, with convolutional neural networks (CNNs) emerging as prominent players [9]. CNNs are object-recognition-focused networks that excel in pattern recognition tasks [10]. Their architecture has been continuously improved to enhance performance [11], and they find applications across various knowledge domains, particularly in image classification [12], [13]. Noteworthy CNN-based architectures include which combines CNNs with long short-term memory networks (CNN-LSTM) for sequential data processing [14], R-CNN region-based networks [15], and the fast R-CNN, which improves detection speed for region-based networks [16]. These advancements in CNN-based architectures have greatly contributed to the progress of face identification techniques based on deep learning.

Deep learning has facilitated the development of various applications in face recognition, including real-time human action recognition, with CNN-based models demonstrating impressive performance [17]–[20]. However, within the current state of the art, there needs to be more comparative evaluation for different CNN architectures, specifically face detection. This work aims to address this gap by comprehensively evaluating 10 CNN-based architectures using transfer learning [21].

By focusing on face detection, this research contributes to a better understanding of the effectiveness and suitability of various CNN models in this specific domain. Among the applications that this comparative analysis allows is the development of access control systems by user recognition, among others. The article presents the methodology employed based on the use of convolutional networks by transfer of learning.

Next, the methods and materials are presented, exposing the database and architectures to be evaluated. The models compared were AlexNet, VGG-16, and VGG-19, GoogLeNet, Inception V3, ResNet-18, ResNet-50 and ResNet-101, and two additional proposed models called shallow CNN and shallow directed acyclic graph with CNN (DAG-CNN). The results section is presented, analyzing the activations of the networks with the best performance, and finally, the conclusions reached are exposed.

## 2. METHOD

A database consisting of three categories is created to carry out the comparison. Two categories are registered users to be recognized (Javier and Robinson). The other category represents a random group of individuals to verify that others are not recognized (Others). For the construction of the database, photos of users' faces are obtained in different positions so that the network can recognize the person, no matter if the face is not completely in front. For the “Others” category, the CelebA [22] database is used, thus obtaining faces with different characteristics, even some similar to the original users to recognize. In total, 3,840 images are used for training, of which 1,940 are in the “Others” category.

The reason for using almost twice as many images within that category as the two users is to give the network more possible characteristics of the unregistered subjects, to avoid that if a person has similar traits, the network can know that the subject is neither of the original. On the other hand, for the validation of the networks, 525 images are used, distributed in 75 images for each of the users and 375 for the category of “Others,” in order to verify that, although there are many different users, the networks are capable of discriminating against them. In Figure 1, it is possible to see samples of the images for each category. The size of the images varies according to the neural network to be used since not all of them have the standard size of 224×224 pixels.

Eight of the most well-known ones are selected to compare the users' facial recognition capacity between different models of CNNs. These models are AlexNet [23], the two versions of the visual geometry group (VGG) model (16 and 19) [24], GoogLeNet [25], Inception V3 [26], and the ResNet models (18, 50 and 101) [27]. It is also proposed to implement two additional basic models to verify if low-depth architectures can maintain a level of recognition as good as the pre-trained models.

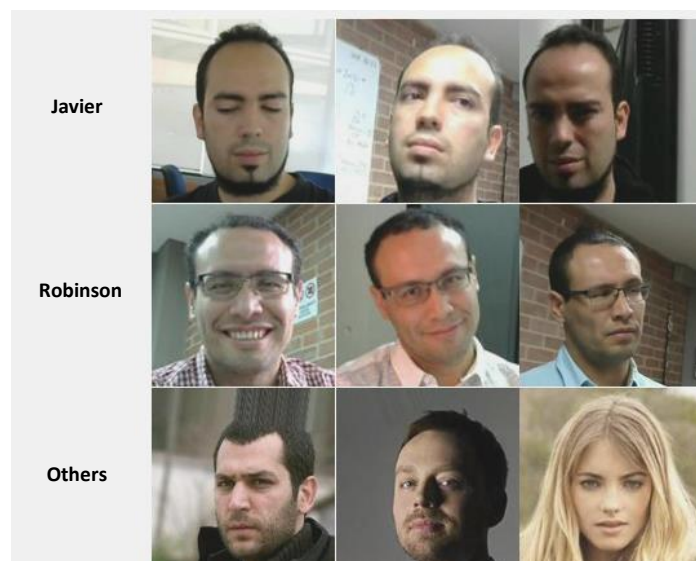


Figure 1. Examples of images used in the database

The two proposed architectures are basic CNN consisting of convolution blocks, where the first one is a sequential network named shallow CNN since it is less deep than its counterparts (apart from AlexNet). The second architecture comprises two branches, one with convolution filters of size  $3 \times 3$ , and another with convolution filters of size  $5 \times 5$ , to learn different patterns of faces. The latter is called shallow DAG-CNN because of its different paths and depth, which remains similar to the previous one, although it has a total of twelve convolution layers. A general diagram of the two architectures can be seen in Figure 2, where S refers to the filter stride, P to the padding used, and the last value represents the number of filters used in that layer. The weights of the two proposed networks were initialized using the He method [28].

The same training parameters were set for all the networks, even for the two proposed architectures, to avoid giving one network a greater advantage than another. For the pre-trained models, mixed transfer learning is performed, i.e., the weights of the first convolution layers are frozen and used as feature extractors while the rest of the layers are fine-tuned. The parameters are as follows: learning rate of  $10^{-3}$  with a reduction factor of 0.5 every four epochs; training will be done for eight epochs, with a mini-batch size of 8 per iteration. These parameters are selected because the models were mostly trained with a learning rate of 0.1, and their weights are expected to not vary greatly from the initial ones. Similarly, it is optional to train for many epochs to avoid over-adjustment. As for the classification section, its learning rate is multiplied by a factor of 10 since, at this stage, the network has not had initial learning, so its rate must be higher for its learning curve to be greater.

During the training, three of the networks had problems with their learning: the AlexNet and VGG models, where their gradient tended to increase abruptly, preventing the network from finishing the training. For this reason, for these models, it was decided to carry out a transfer learning with complete fine-tuning in all its layers. Being pre-entrained and deep networks, it is the generalized characteristics of the architectures that extend the gradient loss.

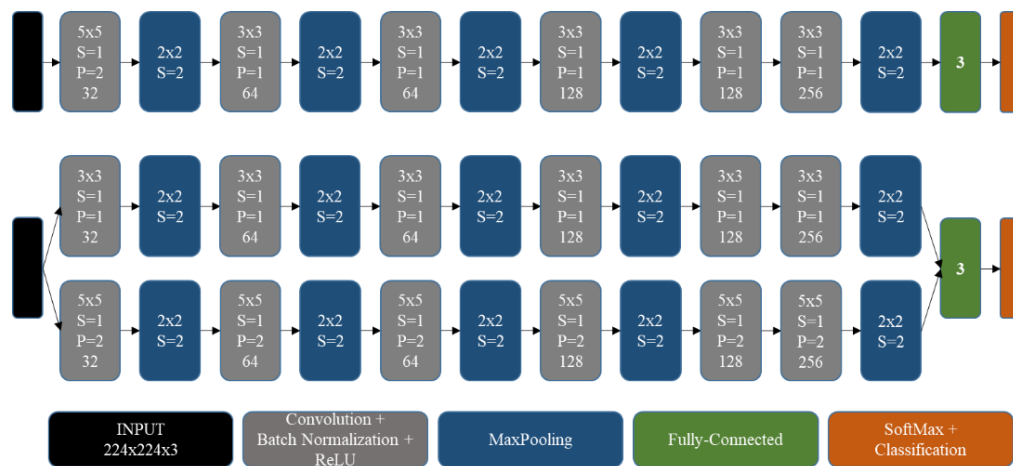


Figure 2. Proposed shallow CNN

### 3. RESULTS AND DISCUSSION

For the first performance evaluation, all networks were compared, as shown in Figure 3, during their training and tested with the validation set, thus obtaining the behavior of the network accuracy in Figure 3(a). In this, the networks with the worst behavior were AlexNet and VGG. Although their losses were reduced in Figure 3(b), these nets fell into overfitting early, remaining below 75% accuracy. On the other hand, the rest of the networks achieved an accuracy above 95%, with shallow DAG-CNN and ResNet-18 as the two slowest learning networks. The fastest models were the GoogLeNet, the ResNet-50, and the ResNet-101, achieving more than 98% accuracy in their first epoch.

The GoogLeNet and ResNet-101 models were the ones that obtained the best recognition performance, managing to discriminate without errors all the images. It is also possible to observe how the two shallow type networks obtained results above 98%; even the DAG-CNN achieved the same result as the Inception V3 without having previous learning and fewer layers than the Inception. AlexNet and the VGG maintained a low level of recognition because they could recognize either of the two users, i.e., the two registered users were recognized as one user, despite correctly classifying the category of “Others”. Table 1 shows the accuracies obtained with each of the trained architectures.

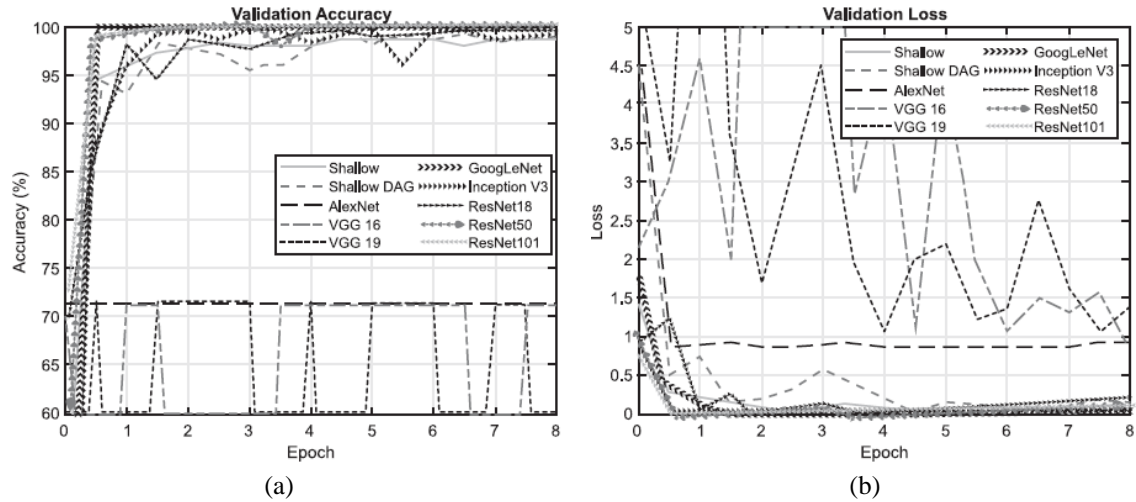


Figure 3. The validation set is used in (a) network accuracy and (b) network loss during training

Table 1. Comparison of user’s face recognition results

Network	Accuracy [%]	Network	Accuracy [%]
Shallow CNN	98.67	GoogLeNet	100
Shallow DAG-CNN	99.05	Inception V3	99.05
AlexNet	71.43	ResNet-18	99.81
VGG-16	71.43	Resnet-50	99.81
VGG-19	14.29	ResNet-101	100

In order to enhance the testing and comparison of the models, it was proposed to verify their operation in real-time, adding a higher level of difficulty in recognition by changing the style of one of the users, that is, the beard and hairstyle. Most photos of the user “Javier” are presented as the examples shown in Figure 1, where he has hair and no facial hair. For real-time testing, the user uses full beard and hair removal to verify if the networks can recognize him. Face detection is performed using the Viola-Jones algorithm [29], with which the bounding box is cropped and then sent to the neural network.

Each neural network was tested with the user “Javier” video sequence. Figure 4 shows a frame taken from the sequence, where the category in which each model classified the user's face is displayed. The AlexNet and VGG models maintained constant user recognition in other categories. On the other hand, the shallow and ResNet-18 types, although capable of recognizing the user, maintained a constant variation of categories, repeatedly classifying the user as unknown, even if the position of the face was frontal, as can be seen in the figure. As for the deeper networks, they managed to maintain an accurate recognition in most of the video, with few category changes, without confusing him with the other user or classifying him as unknown.



Figure 4. Tests performed in real-time



With the models that performed better in the real-time tests, i.e., those with less variation in the user's classification, another test is done, where the user makes slight rotations of the face, to check if the networks could maintain an accurate recognition. In this test, only two networks could maintain a correct classification, the Inception V3 and the ResNet-101, as seen in Figure 5. The percentages of success are respectively 92% and 76%, evidencing the better performance of the Inception architecture used.

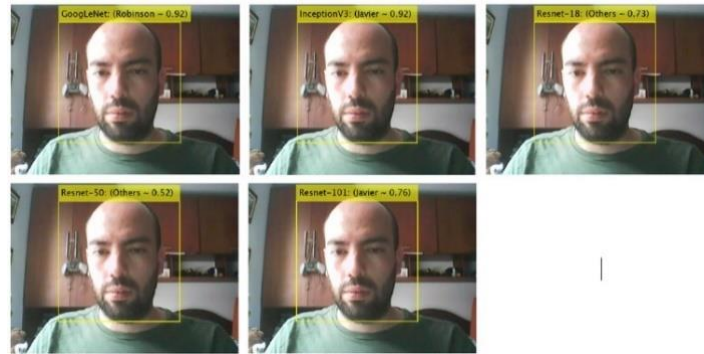


Figure 5. Face rotation tests

The capability of these two networks (the Inception and the ResNet-101) is because, thanks to a large number of convolution blocks, they can learn specific patterns of the user's face, helping them to improve their recognition so that there is a change of style of the user. While architectures with less depth can recognize the user adequately, if a change is made that is not contemplated in the learning set, they will not be able to recognize the user because the learned characteristics will be more general. This factor is shown in Figure 6, where the first layer activations of the shallow DAG-CNN in the  $3 \times 3$  filter-branch Figure 6(a) and the ResNet-101 Figure 6(b) are obtained from an image of the validation set (top) and an image of the test video (bottom).

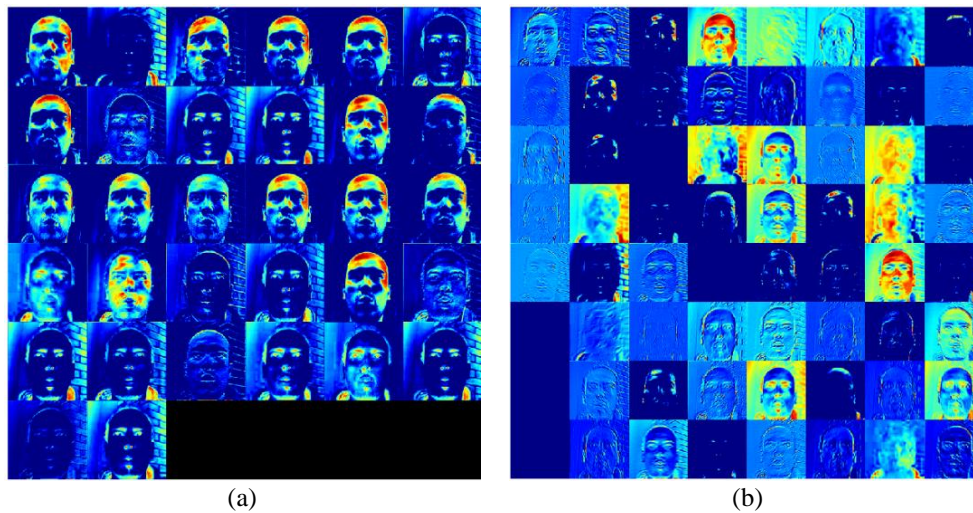


Figure 6. Activations of the first layer in (a) shallow DAG-CNN and (b) ResNet-101 architectures

In shallow architecture, the network focuses on general face patterns, such as eyes, nostrils, and specific face parts. However, these are repeated through most filters without having variations of other user characteristics, as seen in the first image of activations. It can also distinguish shapes and edges, but only in slight sections of the face (such as the shape of the eye), without taking into account, for example, the general shape of the person's skull. Another pattern in their learning is found mainly only in the forehead, repeated by several filters, making a not very specific feature of the user to cover many activations.

As for ResNet-101, although it also focuses its learning on these parts of the face, it does it in a more detailed way, without generalizing or joining several sections at the same time in one filter. But only aiming to discriminate certain patterns, achieving a better distribution of what each filter has learned. For instance, some filters learned the shape of the head, others the location of the eyes, and the nose, apart from its nostrils. Switching to its application in the real-time test further enhances each of the features learned by each network, whereas the shallow mostly keeps its activations on the whole face without discriminating more generally about the characteristics of the user. Meanwhile, ResNet can even highlight the shape of the user's face and head and the location of the ears.

#### 4. CONCLUSION

In this work, a comparison of different CNN models for facial recognition was made to verify the performance of each one. With these comparisons, it was demonstrated that the best networks for this application were GoogLeNet and ResNet-101, which managed to recognize the two users without error correctly and to discriminate against all subjects not belonging to the database. However, the shallow networks without pre-training, such as the shallow CNN and the DAG-CNN, could obtain high performance, even matching the capacity of the Inception V3.

After performing a style change, an additional test was added to recognize one of the users. With this, it was found that the two networks with the greatest capability to withstand drastic changes in certain user characteristics are the Inception V3 and the ResNet-101, which have a greater capacity to learn detailed user features due to their depth. They managed to maintain constant recognition of the subject, even when performing face rotations. This robustness was demonstrated using the layer activations, comparing the learning of one of these against one of little depth, evidencing that these networks could learn more detailed patterns, allowing them to discriminate characteristic features of the user.

#### ACKNOWLEDGMENTS

The authors are grateful to Universidad Militar Nueva Granada for the funding of this Project and Universidad de los Llanos for all help to participate.





#### REFERENCES

- [1] W. Niu, Y. Zhao, Z. Yu, Y. Liu, and Y. Gong, "Research on a face recognition algorithm based on 3D face data and 2D face image matching," *Journal of Visual Communication and Image Representation*, vol. 91, Mar. 2023, doi: 10.1016/j.jvcir.2023.103757.
- [2] R. Rameswari, S. N. Kumar, M. A. Ananth, and C. Deepak, "Automated access control system using face recognition," *Materials Today: Proceedings*, vol. 45, pp. 1251–1256, 2021, doi: 10.1016/j.matpr.2020.04.664.
- [3] S. M. Bah and F. Ming, "An improved face recognition algorithm and its application in attendance management system," *Array*, vol. 5, Mar. 2020, doi: 10.1016/j.array.2019.100014.
- [4] C. Y. J. Liu and C. Wilkinson, "Image conditions for machine-based face recognition of juvenile faces," *Science & Justice*, vol. 60, no. 1, pp. 43–52, Jan. 2020, doi: 10.1016/j.scijus.2019.10.001.
- [5] C. Zhao, X. Li, and Y. Dong, "Learning blur invariant binary descriptor for face recognition," *Neurocomputing*, vol. 404, pp. 34–40, Sep. 2020, doi: 10.1016/j.neucom.2020.04.082.
- [6] A. Shah, B. Ali, M. Habib, J. Frnda, I. Ullah, and M. S. Anwar, "An ensemble face recognition mechanism based on three-way decisions," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 4, pp. 196–208, Apr. 2023, doi: 10.1016/j.jksuci.2023.03.016.
- [7] F. Tang *et al.*, "An end-to-end face recognition method with alignment learning," *Optik*, vol. 205, Mar. 2020, doi: 10.1016/j.ijleo.2020.164238.
- [8] D. Manju and V. Radha, "A novel approach for pose invariant face recognition in surveillance videos," *Procedia Computer Science*, vol. 167, pp. 890–899, 2020, doi: 10.1016/j.procs.2020.03.428.
- [9] J. Yuan *et al.*, "Gated CNN: Integrating multi-scale feature layers for object detection," *Pattern Recognition*, vol. 105, Sep. 2020, doi: 10.1016/j.patcog.2019.107131.
- [10] I. Rafegas, M. Vanrell, L. A. Alexandre, and G. Arias, "Understanding trained CNNs by indexing neuron selectivity," *Pattern Recognition Letters*, vol. 136, pp. 318–325, Aug. 2020, doi: 10.1016/j.patrec.2019.10.013.
- [11] A. M. S. Aradhya, A. Ashfahani, F. Angelina, M. Pratama, R. F. de Mello, and S. Sundaram, "Autonomous CNN (AutoCNN): A data-driven approach to network architecture determination," *Information Sciences*, vol. 607, pp. 638–653, Aug. 2022, doi: 10.1016/j.ins.2022.05.100.
- [12] J. Qin, W. Pan, X. Xiang, Y. Tan, and G. Hou, "A biological image classification method based on improved CNN," *Ecological Informatics*, vol. 58, Jul. 2020, doi: 10.1016/j.ecoinf.2020.101093.
- [13] Y. Li *et al.*, "Robust detection for network intrusion of industrial IoT based on multi-CNN fusion," *Measurement*, vol. 154, Mar. 2020, doi: 10.1016/j.measurement.2019.107450.
- [14] Q. Fu, C. Wang, and X. Han, "A CNN-LSTM network with attention approach for learning universal sentence representation in embedded system," *Microprocessors and Microsystems*, vol. 74, Apr. 2020, doi: 10.1016/j.micpro.2020.103051.
- [15] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, "Instance segmentation of apple flowers using the improved mask R-CNN model," *Biosystems Engineering*, vol. 193, pp. 264–278, May 2020, doi: 10.1016/j.biosystemseng.2020.03.008.
- [16] G. Rajeshkumar *et al.*, "Smart office automation via faster R-CNN based face recognition and internet of things," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100719.
- [17] A. Budiman, R. A. Yaputera, S. Achmad, and A. Kurniawan, "Student attendance with face recognition (LBPH or CNN): systematic





- literature review,” *Procedia Computer Science*, vol. 216, pp. 31–38, 2023, doi: 10.1016/j.procs.2022.12.108.
- [18] F. Zhao, J. Li, L. Zhang, Z. Li, and S.-G. Na, “Multi-view face recognition using deep neural networks,” *Future Generation Computer Systems*, vol. 111, pp. 375–380, Oct. 2020, doi: 10.1016/j.future.2020.05.002.
- [19] S. R. Mishra, T. K. Mishra, G. Sanyal, A. Sarkar, and S. C. Satapathy, “Real time human action recognition using triggered frame extraction and a typical CNN heuristic,” *Pattern Recognition Letters*, vol. 135, pp. 329–336, Jul. 2020, doi: 10.1016/j.patrec.2020.04.031.
- [20] K. B. Pranav and J. Manikandan, “Design and evaluation of a real-time face recognition system using convolutional neural networks,” *Procedia Computer Science*, vol. 171, pp. 1651–1659, 2020, doi: 10.1016/j.procs.2020.04.177.
- [21] J. Lin, L. Zhao, Q. Wang, R. Ward, and Z. J. Wang, “DT-LET: deep transfer learning by exploring where to transfer,” *Neurocomputing*, vol. 390, pp. 99–107, May 2020, doi: 10.1016/j.neucom.2020.01.042.
- [22] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” *Prepr. arXiv.1411.7766*, Nov. 2014.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Prepr. arXiv.1409.1556*, Sep. 2014.
- [25] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: surpassing human-level performance on ImageNet classification,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.
- [29] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004, doi: 10.1023/B:VISI.0000013087.49260.fb.

## BIOGRAPHIES OF AUTHORS







**Javier Orlando Pinzón Arenas**     was born in Socorro-Santander, Colombia, in 1990. He received his degree in mechatronics engineering (cum laude) and specialization in Engineering Project Management at Universidad Militar Nueva Granada-UMNG in 2013 and 2016, respectively. He has experience in the areas of automation, electronic control, and machine learning. Currently, he has a degree in mechatronics engineering and is working as a research assistant at the UMNG with an emphasis on robotics and machine learning. He can be contacted at est.javier.pinzon@unimilitar.edu.co.



**Robinson Jiménez-Moreno**     is an electronic engineer who graduated from Universidad Distrital Francisco José de Caldas in 2002. He received an M.Sc. in engineering from Universidad Nacional de Colombia in 2012 and a Ph.D. in engineering at Universidad Distrital Francisco José de Caldas in 2018. He is currently working as an assistant professor at Universidad Militar Nueva Granada and his research focuses on the use of convolutional neural networks for object recognition and image processing for robotic applications such as human-machine interaction. He can be contacted at robinson.jimenez@unimilitar.edu.co. His profile can be found at ResearchGate <https://www.researchgate.net/profile/Robinson-Moreno-2>. RedDOLAC <https://reddolac.org/profile/RobinsonJimenezMoreno>.



**Javier Eduardo Martínez Baquero**     is an electronic engineer who graduated from Universidad de los Llanos in 2002, a postgraduate in electronic instrumentation from Universidad Santo Tomás in 2004, a postgraduate in instrumentation and industrial control at Universidad de los Llanos in 2020, and an M.Sc. in educative technology and innovative media for education at Universidad Autónoma de Bucaramanga in 2013. He is currently working as an associate professor at Universidad de los Llanos and his research focuses on instrumentation, automation, control, and renewable energies. He can be contacted at jmartinez@unillanos.edu.co. His profile can be found at ResearchGate <https://www.researchgate.net/profile/Javier-Martinez-Baquero> and RedDOLAC [https://reddolac.org/profile/JavierEduardoMartinezBaquero?xg\\_source=activity](https://reddolac.org/profile/JavierEduardoMartinezBaquero?xg_source=activity).