Small Face Detection Using Deep Learning on Surveillance Videos

Rolando J. Cárdenas, Cesar A. Beltrán, and Juan C. Gutiérrez

Abstract—**Face detection is one of the essential tasks widely studied in the field of Computer Vision. Several authors have developed different techniques to improve the face detection in images, but these are limited in their application on videos and more if they present low resolution. In this study, we propose a new model for face detection in low-resolution videos based on the morphology of the upper body of people, and the use of Deep Learning (CNN). Our results show an average of 39% accuracy over the Caviar dataset and 32% in the UCSP dataset. Compared with other techniques, our results are greater due they only reach 1% of accuracy.**

Index Terms—**Deep learning, face detection, low resolution, video.**

I. Introduction

Video cameras used in surveillance are becoming more important in companies, cities, and small businesses that seek to keep their goods safe from the possibility of theft, robbery, or illicit activities of their staff [1]. Currently, cities are more frequently relying on surveillance cameras to prevent or alert criminal acts and to delve into the causative reasons for traffic concerns or car accidents [2]–[4].

Surveillance videos also help in face detection. It is carried out by analyzing the physical traits of people's faces and located where it is. However, the current face recognition systems have reached a certain level of maturity, but the development in videos remains limited due to the conditions presented in outdoor environments. For example, the face detection process in images obtained from outdoor videos has variations caused by changing illumination conditions that cannot be easily controlled. In addition, the partial or total occlusion with others objects and the view angle due to the camera position or low-resolution sensors of the acquired images make face recognition all the more difficult. All characteristics of these frames make more difficult apply techniques such as face detection or recognition which were initially designed for pictures in semi-controlled environments, such as a laboratory or any indoor environment [2], [5]–[14].

In this study, we explore how Deep Learning affects the process of face detection in low-resolution surveillance videos, because is common to see surveillance cameras installed in higher locations for monitoring a full urban zone [15], and this is where the previous works on face detection have limitations due to the size of faces in which they performed. Experiments performed using our proposed model with Deep Learning technique have shown significant improvements in the accuracy of face detection in low-resolution videos.

This article is organized as follows: Section II presents previous works related to face detection, Section III describes the methodology proposed for face detection using our Deep Learning technique, Section IV summarizes the results obtained from our experiments, and Section V provides study conclusion and suggests the type of future work that needs to be conducted.

II. Previous Works

Herrmann *et al.* [15] presented a process of face detection in low-resolution videos (faces with a size of <100 pixels) using the Viola-Jones face detector, a commonly used method for real-time face detection [16], trained on low-resolution face images. They reported achieving 98.7% and 0.27% accuracy rates in the detection of faces from image sizes of >20 and <14 pixels, respectively.

Qiang *et al.* [17] proposed a face detection model that used a head and shoulder cascade detector and Histograms of Oriented Gradients (HOG) [18] for image detection. They reported achieving 83.9% accuracy in the detection of faces from an image size of 64 × 80 pixels.

Mutneja *et al.* [19] proposed a face detection algorithm for low-resolution videos based on frame differences, integral images, and Haar cascade classifiers. This approach focuses on speed detection based on low-processing techniques. The authors reported a 98% accuracy rate in face detection, but over a close range, using this approach.

Low-resolution videos are known to produce the blur effect, comparable to low-resolution images. Zhang-Xiang *et al.* [20] proposed a new approach for blurred face recognition. They were able to achieve 95% accuracy in face recognition using images as small as 128 × 128 pixels.

Zhang *et al.* [21] proposed a novel Densely Connected Face Proposal Network. The architecture consists of two units: Rapidly Digested Convolutional Layers designed to reduce the spatial size of images and the Densely Connected Convolutional Layers designed to enrich the receptive field of the last convolutional layer. Although the model achieved 98.49% accuracy when tested on AFW datasets, it could only detect faces with a size of >40 pixels.

Triantafyllidou *et al.* [22] proposed a lightweight deep Convolutional Neural Network for face detection trained with a progressive, positive method that allows for

Manuscript received October 22, 2018; revised February 22, 2019.

Rolando J. Cárdenas and Juan C. Gutiérrez are with National University of San Agustin – UNSA, Arequipa, Peru (e-mail: jecardenast200@gmail.com, jcgutierrezc@gmail.com).

Cesar A. Beltrán is Pontificia Universidad Católica del Perú – PUCP, Artificial Intelligence Research Group (IA-PUCP), Lima, Peru (e-mail: cbeltran@pucp.pe).

doi: 10.18178/ijmlc.2019.9.2.785

identification of facial parts (e.g., eyes, mouth, and nose). When applied on the FDDB datasets, this model was able to achieve a recall rate of 92.6%.

Sun *et al*. [23] improved the state-of-the-art faster RCNN framework by combining feature concatenation, hard detrimental mining, multiscale training, and calibration of critical parameters. This model achieved 80% accuracy.

Sawat and Hegadi [24] proposed a model using deep features extracted by deep CNN and the classification by Cubic Support Vector Machine. The application of the model over IJB-A database achieved an accuracy rate of 98.2%.

Yang *et al*. [25] proposed a deep convolutional network that achieves a recall rate of 90.99% when applied on the FDDB database. However, the main drawback of this model is that it uses physical features to locate faces in the images.

As discussed in the previous paragraphs, some models were developed to improve the accuracy of face recognition in low-resolution images. As can be seen from their results, these systems tend to perform well for <100-pixel images.

Furthermore, although Deep Learning Networks have been developed, they perform well only for well-illuminated images or require extensive hardware for video processing. Deep Learning Networks, therefore, are not designed to perform well for processing low-resolution videos.

In this study, we propose a new face detection model dedicated for processing low-resolution videos with a new low-resolution dataset for Deep Learning training, which works effectively on low-resolution videos and on any face scale (regardless of the person distance to the camera).

III. Methodology

We propose a new face detection model enabled with Deep Learning. The architecture of the model is divided into five stages. Fig. 1 shows the complete pipeline of the proposed model.

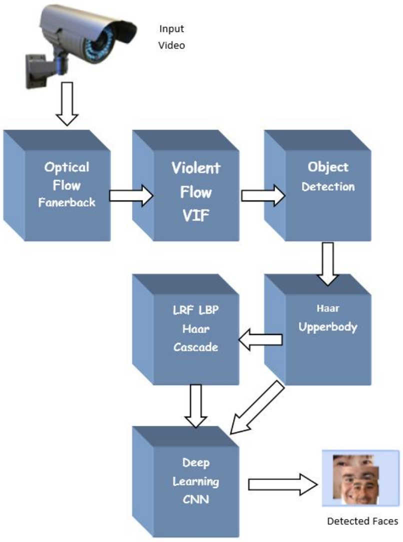

Fig. 1. Proposed model with the five stages of the methodology, starting with the motion object segmentation (optical flow), person descriptor identification (violent flow), detection of motion person areas, the upper body and face detection, and at the end the Deep Learning stage.

The proposed model is designed to be applied only for videos due to motion segmentation through the optical flow algorithm. We used the Gunnar Farneb äck optical flow algorithm [26] because it is a dense optical flow algorithm that helps calculate the movement of all pixels of the frame. A detailed explanation of the optical flow algorithm to Haar modules can be found elsewhere [27].

A. Deep Learning – Convolutional Neural Network

Fig. 2 illustrates the complete architecture of the Convolutional Neural Network (CNN) used in the last stage of the proposed model.

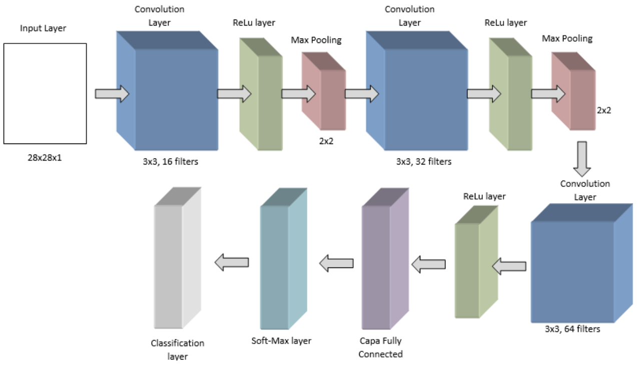

Fig. 2. Convolutional Neural Network architecture.

The CNN is composed of 12 layers. An input layer with the size of 28 × 28 pixels. Three convolutional layers with 16, 32, and 64 filters and size of 3 × 3. Three Rectified Linear Unit layers (ReLu). Two max-pooling layers with stride 2. A fully connected layer. Next, a SoftMax Layer, and finally, the classification layer with two class (Face and background).

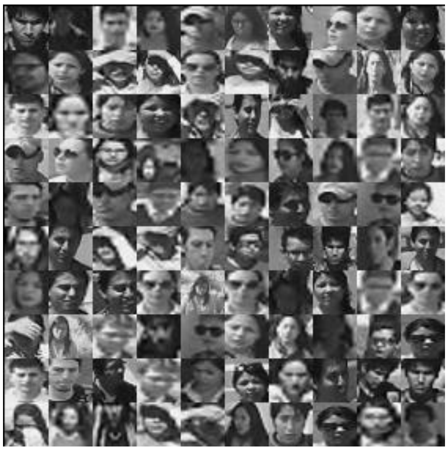

Fig. 3. Example of true low-resolution faces used in the Deep Learning training.

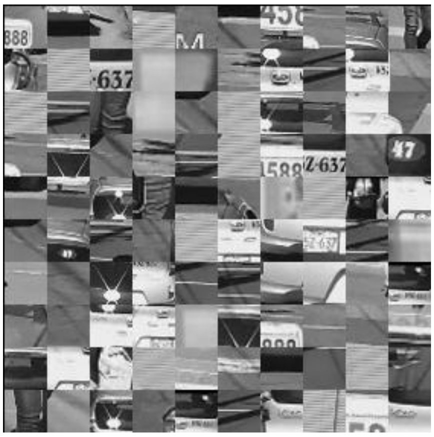

Fig. 4. Example of false low-resolution faces used in deep learning training.

This CNN was trained with small face images as same as low-resolution faces (LBP). Fig. 3 and Fig. 4 show samples of pictures used in the training process, positive and negative sets, respectively.

IV. Experiments

For our study, the Caviar video database [29] and the UCSP video database were used [30]. First, we tested the behavior of the Deep Learning model with several images in the training step; second, the best models were tested in both databases, and the obtained results were summarized.

A. Deep Learning Training

The Deep Learning model was trained on both databases separately, thereby producing two different models, each trained on a different database.

Face images were extracted manually from the Caviar and UCSP video databases. For each dataset, 70% of the pictures were used for training and the rest of the images for testing. Table I shows the results obtained for the training and testing datasets. In addition, we tested each model with the other database (Caviar network with UCSP database, and vice versa). The UCSP Network achieved the best results with an accuracy rate of 92.75%.

TABLE I: Training Dataset

	Caviar Dataset (3000 Images)	UCSP Dataset (2200 Images)
Caviar1 network	92.06	75.70
UCSP1 network	67.52	**92.75**

The second experiment was conducted on both the datasets, but with data augmentation. On each image of the datasets, we applied rotations such as 10 º clockwise, 10 º anti-clockwise, and mirror image.

The total number of images obtained after applying the rotations was 24,000 for Caviar dataset and 8,800 for UCSP dataset. We trained two Deep Neural Networks by using 70% of the images from each dataset for training and the remaining images for testing. Furthermore, we tested each model with the other database. Table II shows the results of the experiment in which the UCSP network presents the best score, reaching 95.34% accuracy.

Considering these results, we used the best two networks for the posterior experiments.

TABLE II: Data Augmentation Training Dataset

	Caviar Dataset (24,000 Images)	UCSP Dataset (8,800 Images)
Caviar2 network	93.63	73.98
UCSP2 network	66.72	**95.34**

B. Caviar Database

In this experiment, we used the Caviar video dataset [29] to test our proposed strategy with both Deep Learning models.

Fig. 5 shows example frames extracted from the Caviar video dataset. We tried to detect the small-sized faces seen in this database. Additionally, we could see the detailed features such as the number of people and duration of each video (Table III).

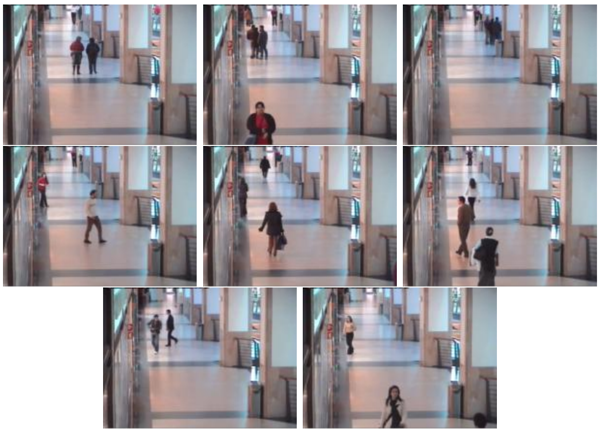

Fig. 5. Example frames extracted from Caviar video dataset. Notice the low resolution of the faces.

TABLE III: Caviar Video Database

No.	Video	Duration (s)	No. People
1	EnterExitCrossingPaths1cor	00:15	5
2	OneLeaveShop1cor	00:11	7
3	OneLeaveShop2cor	00:44	6
4	OneLeaveShopReenter1cor	00:15	4
5	OneLeaveShopReenter2cor	00:22	7
6	OneShopOneWait1cor	00:55	10
7	OneStopEnter2cor	01:49	8
8	OneStopMoveNoEnter1cor	01:06	6

In addition to the results of previous studies [27] that evaluated the OpenCV 2.4.13 library (with Haar Cascades detector [16]), Dlib C++ Library (HOGs [18]), MATLAB R2017a (using their vision.CascadeObjectDetector function based on Haar Cascades), Castrillon Upper Body technique [28], the proposal using Castrillon Upper Body technique [28] alone, and the proposal LRF-LBP, we evaluated the accuracy of face detection using both Deep Learning models.

All these techniques were tested on the eight videos that were selected and previously manually analyzed from the Caviar database.

We applied the F-Score measure (Equation 1) as criteria for evaluation. Table IV shows the results with different approaches.

$$\text{FScore} = 2 * \frac{\text{Precision} \ * \text{Recall}}{\text{Precision} \ + \text{Recall}} \quad (1)$$

TABLE IV: Evaluation of the Overall Performance (F-Score) with Other Techniques

	Average
Dlib c++ Library	00.00
OpenCV Library	00.85
Matlab (Haar)	01.21
Castrillon	21.07
Proposal + Haar Upper Body	46.48
Proposal + Haar Upper Body + LRF-LBP	50.68
Proposal + UCSP1 network	39.50
Proposal + UCSP2 network	**42.86**

The obtained results with both Deep Learning Neural Networks show an average of 39.50% and 42.86%, respectively. These results are lower than we expected, due that Deep Learning rejects some small correct face detections of the previous step. However, these averages contain a lower rate of false positives. Fig. 6 illustrates this behavior. The proposal with the highest accuracy in Table IV shows the highest false positive rate in Fig. 6, so this implementation has a mean of 1.7% of wrong detections per frame.

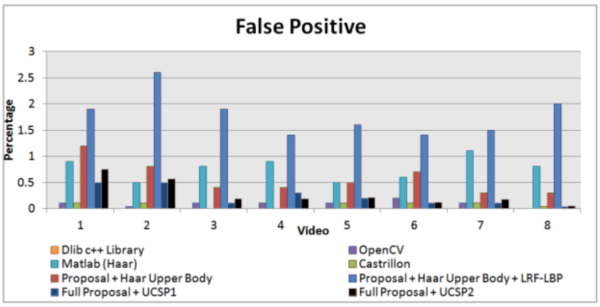

Fig. 6. Bar chart for the overall performance.

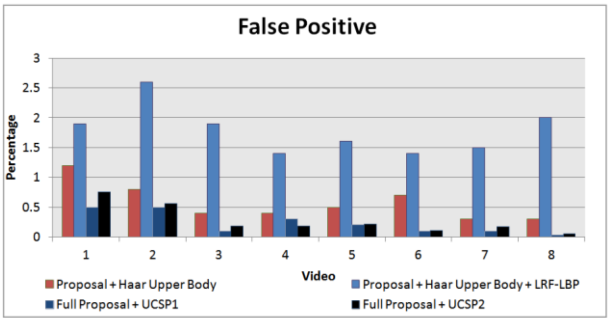

Fig. 7. False positive obtained from the three top overall performance bar charts.

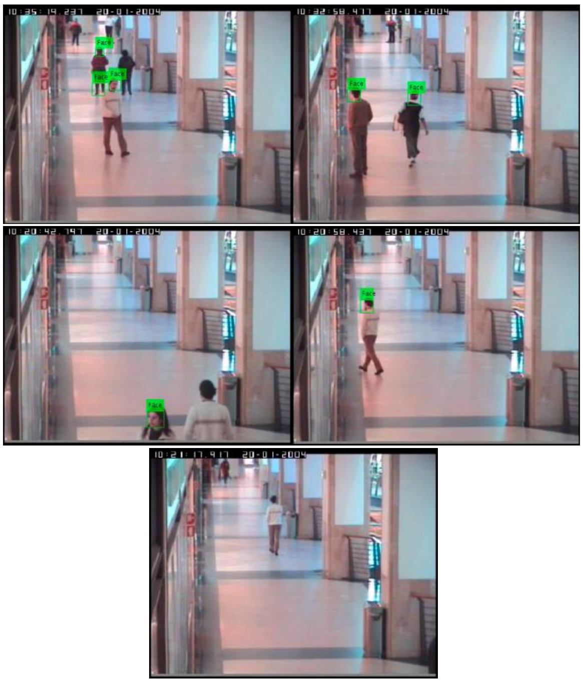

Fig. 8. Obtained results with the proposed model on UCSP2 network.

By contrast, the Deep Leaning models are more accurate in facial recognition despite their lower performance, because of the lower false-positive rate than other techniques with a high accuracy. Comparing the four top overall performance methods (Fig. 7), the two Deep Learning Networks had a lower false-positive rate (UCSP1 network with 0.22% and UCSP2 network with 0.27% of wrong detections per frame) and achieved a mean accuracy of 40%. Therefore, these models are deemed to have the ability to detect small faces in these datasets.

Finally, examples of the results obtained with the UCSP2 network are illustrated in Fig. 8. Of note is the minimum false detection rate.

C. UCSP Database

In this experiment, the UCSP database [30] was used for testing our proposed model with both Deep Learning Networks.

The UCSP Dataset was obtained from a Dahua surveillance video camera with an HD resolution at a rate of 30 fps. The recorded videos correspond to people entering and leaving a laboratory, and they were used by Machacas *et al.* in their work [30]. Fig. 9 shows the sample video frames that were extracted. The changes in illumination in the outdoor and indoor images are obvious.

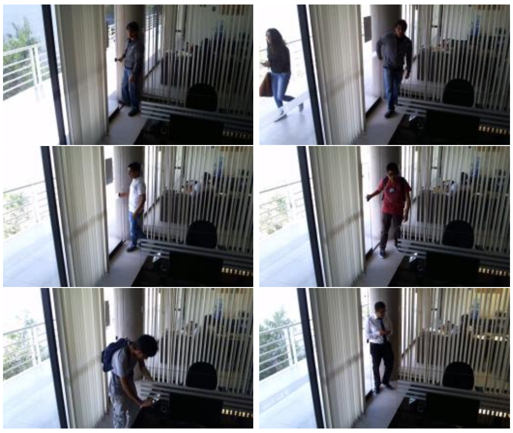

Fig. 9. Sample frames extracted from the UCSP dataset. These videos were reduced to the half to simulate low resolution.

The details of each video (duration of the video in seconds and the number of people that enter and leave the scene) are summarized in Table V.

TABLE V: UCSP VIDEO DATABASE

No.	Video	Duration (s)	No. People
1	Video 1	00:07	1
2	Video 2	00:14	2
3	Video 3	00:10	1
4	Video 4	00:20	4
5	Video 5	00:08	1
6	Video 6	00:10	1

We compared our proposed model with that proposed by Machaca *et al.* [30] for low-resolution videos without any super-resolution or illumination normalization algorithms over the USCP database.

We resized the video frames to half (720 × 480 pixels) to simulate a low-resolution video, although the illumination of the scenes was not modified.

Our proposal without any Deep Learning model achieved an average of 22% accuracy, and that with the UCSP1 network achieved an average of 25% accuracy (Table VI). Furthermore, the proposal with the UCSP2 network achieved a better average of 33%. This result shows how our proposal improves more with the two Deep Learning models and

UCSP2 network obtained the best results, which used additional data for the training step.

Moreover, these experiments did not use techniques such as illumination normalization or super-resolution, and the results including the proposal without any Deep Learning model are more significant than those obtained by the proposal for low-resolution face detection without super-resolution techniques by Machaca *et al*. [30].

TABLE VI: ACCURACY IN REAL SURVEILLANCE VIDEOS (UCSP DATASET)

	Video						Average
	1	2	3	4	5	6	
Without Deep Learning	9.34	35.1	27.9	21.3	11.8	27.2	22.11
UCSP1 network	41.54	19.67	35.83	31.00	0.00	26.7	25.79
UCSP2 network	**45.0**	**31.87**	**39.37**	**41.64**	**0**	**39.6**	**32.91**

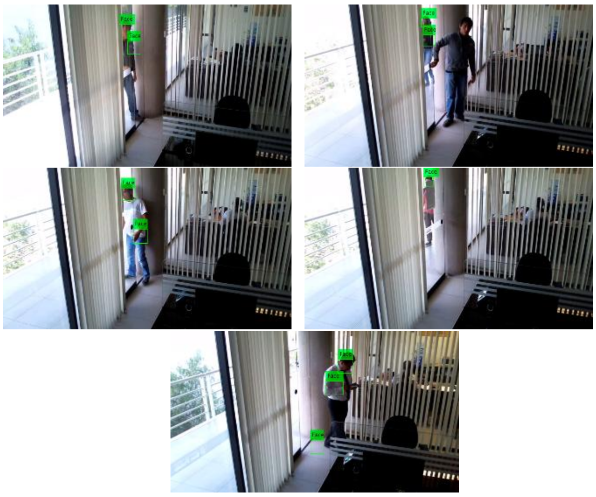

Fig. 10. Results obtained with the proposed model on UCSP2 network.

Results obtained with the proposal using the UCSP2 network are shown in Fig. 10. We highlight the detections obtained from outside because these faces are not easy to detect with another proposal due that are not show completely.

V. CONCLUSION AND FUTURE WORK

In this study, we explored the use of Deep Learning in our proposed model to improve the face recognition rates in low-resolution scenarios. The results showed a significant increase in the accuracy of the proposed model with a low rate of false positives on low-resolution videos.

Our results showed lower accuracy and false-positive rates when the proposed model was used on the Caviar database. Therefore, we can rely more on these networks because the majority of the detections are faces and a small percentage of these detections are incorrect (0.2% of wrong detections per frame).

In the UCSP database, the results showed an improvement of 32% in the accuracy rate. This result is far better because it improves in 10% the obtained results with the proposal without Deep Learning techniques, and we avoid the use of super-resolution algorithms or illumination normalization techniques. Moreover, this improvement is higher than Machaca's detector, so Deep Learning shows an essential increase in the face detection.

In our future work, we will continue analyzing the effects of including more datasets on training and the methods to improve the Deep Learning model by modifying the hyperparameters and fine tuning.

ACKNOWLEDGMENT

This research was supported by CIENCIACTIVA, CONCYTEC, and the National University of San Agustin (UNSA). The authors thank all the professors who collaborated in this research.

REFERENCES

[1] C. L. Devasena, R. Revathi, and M. Hemalatha, "Video surveillance systems, a survey," *IJCSI International Journal of Computer Science Issues*, vol. 8, 2011.

[2] M. A. R. Ahad, "Motion history images for action recognition and understanding," *Ser. Springer Briefs in Computer Science*, Springer-Verlag London, 2013.

[3] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *Proc. 2008 37th IEEE Applied Imagery Pattern Recognition Workshop*, Oct 2008, pp. 1–8.

[4] B. Boufama and M. A. Ali, *Tracking Multiple People in the Context of Video Surveillance*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 581–592.

[5] T. I. Dhamecha, G. Goswami, R. Singh, and M. Vatsa, *On Frame Selection for Video Face Recognition*, Cham: Springer International Publishing, 2016, pp. 279–297.

[6] B. Mandal, R. Y. Lim, P. Dai, M. R. Sayed, L. Li, and J. H. Lim, *Trends in Machine and Human Face Recognition*, Springer International Publishing, 2016, pp. 145–187.

[7] A. Thamizharasi and J. Jayasudha, "A literature survey on various illumination normalization techniques for face recognition with fuzzy k nearest neighbour classifier," *ICTACT Journal on Image & Video Processing*, vol. 5, no. 4, 2015.

[8] Z. Zhang, C. Wang, and Y. Wang, *Video-Based Face Recognition: State of the Art*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–9.

[9] M. Davis, S. Popov, and C. Surlea, *Real-Time Face Recognition from Surveillance Video*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 155–194.

[10] H. Wang, Y. Wang, and Y. Cao, "Video-based face recognition: A survey," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 3, no. 12, 2009.

[11] R. Liu, X. Gao, R. Chu, X. Zhu, and S. Z. Li, *Tracking and Recognition of Multiple Faces at Distances*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 513–522.

[12] M. D. Levine, M. R. Gandhi, and J. Bhattacharyya. (2004). Image normalization for illumination compensation in facial images. Department of Electrical & Computer Engineering & Center for Intelligent Machines, McGill University, Montreal, Canada. [Online]. Available: https://pdfs.semanticscholar.org/4657/d87aebd652a5920ed255dca993353575f441.pdf

[13] A. K. Makhtar, H. Yussof, H. Al-Assadi, L. C. Yee, M. Emadi, M. Khalid, R. Yusof, and F. Navabifar, "International symposium on robotics and intelligent sensors 2012 (IRIS 2012) illumination normalization using 2d wavelet," *Procedia Engineering*, vol. 41, pp. 854–859, 2012.

[14] D. Fermi, S. S. Kartha *et al*., "A survey on different face detection algorithms in image processing," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 6, pp. 151–156, 2017.

[15] C. Herrmann, C. Qu, and J. Beyerer, "Low-resolution video face recognition with face normalization and feature adaptation," in *Proc. 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Oct. 2015, pp. 89–94.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001,vol. 1, pp. I–511.

[17] Q. Liu, W. Zhang, H. Li, and K. N. Ngan, "Hybrid human detection and recognition in surveillance," *Neurocomput.*, vol. 194, no. C, pp. 10–23, Jun. 2016.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. the 2005 IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, Washington, DC, USA: IEEE Computer Society, 2005, pp. 886–893.

[19] V. Mutneja and S. Singh, "Face detection and extraction from low resolution surveillance video using motion segmentation," *International Journal on Computer Science and Engineering*, vol. 9, pp. 275–282, 2017.

[20] Z.-X. Feng, Y. Yuan, and J.-H. Lai, *Learning Blur Invariant Face Descriptors for Face Verification Under Realistic Environment*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 355–365.

[21] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Detecting face with densely connected face proposal network," *Biometric Recognition*, Cham. Springer International Publishing, pp. 3–12, 2017.

[22] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big Data Research*, vol. 11, pp. 65–76, 2018.

[23] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *CoRR*, 2017.

[24] D. D. Sawat and R. S. Hegadi, "Unconstrained face detection: A deep learning and machine learning combined approach," *CSI Transactions on ICT*, vol. 5, no. 2, pp. 195–199, 2017.

[25] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015b, pp. 3676–3684.

[26] G. Farnebäck, *Two-Frame Motion Estimation Based on Polynomial Expansion*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.

[27] T. R. J. Cárdenas, C. A. B. Castañón, and J. C. G. Cáceres, "Face detection on real low-resolution surveillance videos," in *Proc. the 2nd International Conference on Compute and Data Analysis*, New York, NY, USA, 2018, pp. 52-59.

[28] M. C. Santana, J. L. Navarro, O. D. Suarez, J. I. Gonzalez, and A. F. Martel, *Multiple Face Detection at Different Resolutions for Perceptual User Interfaces*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 445–452.

[29] Caviar test case scenarios. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[30] V. M. Arceda, K. F. Fabian, and J. Gutierrez, "Real time violence detection in video," *IET Conference Proceedings*, January 2016.

Rolando Jesus Cardenas T. is from Arequipa, Peru. He got a bachelor in computer science from the National University of San Agustin of Arequipa in 2012, and current he is a magister student on information technology at National University of San Agustin, Arequipa, Peru. He is also a researcher in computer vision, image processing, video processing, and machine learning.

César A. Beltrán Castañón is from Lima, Perú. He is a profesor and senior reseracher in Pontificia Universidad Católica del Perú, Dpto. de Ingeniería, Seccion Ing. Informática. He was a post-doctoral in Texas A&M University in 2016, bioinformatic doctor and magister on computer science at USP Sao Paulo, Brazil. He was the vice-president of the IEEE Computer Society Peru from 2015 to 2018. He is the president of the Peruvian Association of Pattern Recognition. He is also the scientific leader of the Center for Scientific Innovation and Technological Development in Computer Sciences of the PUCP, scientific leader of the Artificial Intelligence Group of the PUCP (IA-PUCP). His research areas are in machine learning, data analytics, deep learning, computational vision, image processing, content image retrieval algorithms, bioinformatics, high-performance computing.

Juan Carlos Gutiérrez Cáceres is from Arequipa, Peru. He was a doctor in computer science at the National University of San Agustín (Peru). He was a master in computer science and computational mathematics at the Institute of Mathematical and Computer Sciences (ICMC) of the University of São Paulo (Brazil). He is a founding member of the Peruvian Computer Society (SPC) Peru. Currently he holds the position of director of the Professional School of Computer Science of the National University of San Agustín, Arequipa, Peru. He is a professor in the Professional Program of Computer Engineering at the San Pablo Catholic University. His areas of interest are in complex networks patterns recognition, neural networks, nonlinear dynamic systems, image processing.