Video Analysis with Faces using Harris Detector and Correlation

Rodolfo Romero Herrera[1]
Departamento de Ciencias e Ingeniería de la Computación
Instituto Politécnico Nacional-ESCOM, Ciudad de México, México

Francisco Gallegos Funes[2]
Sección de Estudios de Posgrado e Investigación, Instituto Politécnico Nacional- ESIME, Ciudad de México, México

José Elias Romero Martínez[3]
Sección de Estudios de Posgrado e Investigación, Instituto Politécnico Nacional, Ciudad de México, México

***Abstract*—A procedure is presented to detect changes in a video sequence based on the Viola & Jones Method to obtain images of faces of persons; videos are taken of the web. The software allows to obtain images or frames separated from nose, mouth, and eyes but the case of the eyes is taken as an example. Change detection is done by using correlation and the Harris detector. The correlation results allow us to recognize changes in position in the person or movement of the camera if the individual remains fixed and vice versa. It is possible to analyze a part of the face thanks to the Harris detector; It is possible with detected reference points to recognize very small changes in the video sequence of a particular organ; such as the human eye, even when the image is of poor quality as is the case of videos downloaded from the Internet or taken with a low-resolution camera.**

***Keywords*—Harris detect; Viola and Jones; Harris detector; correlation; video**

I. Introduction

Face analysis is a method used in several applications such as cell phone security systems, to detect stress levels, in gender recognition, etc. [1] [2] [3]; Another topic of interest is undoubtedly the follow-up of the face and its analysis in video sequences [4] [5]; and it is because there is more information in a temporary space than if only one image is analyzed; since, without a doubt, the dependence between one image and another on a video sequence reveals information of interest; However, the problem is complicated, if the video is of poor quality as are internet videos.

The relationship between frames or consecutive images can give us, for example, the recognition of an important change pattern such as the state of emotion between sequences or a significant change due to an unexpected event [6] [7]. This can be measured by the correlation in contours or landmark [8] [9]. For example, if you want to measure the duration of an emotion or the unexpected change of it [10]. Thus applying some processing to a bad video will result in a bad analysis. For this reason, it is proposed to apply such analysis to landmarks, in order to obtain results for these cases, no matter the poor of the video.

In this investigation, the Viola & Jones method is used for the location of the face in video sequences and the segmentation in RGB for its follow-up [11] [12]. Subsequently, a probabilistic analysis is used to see the relationship between one image and another and detect changes [13]. It is feasible to separate parts of the face in the mouth, nose, and eyes, for the analysis of their individual sequence. It was chosen to process the eyes because it involves greater difficulties; because processing smaller images or frames is more difficult, Due to the fact, they are taken from bodies or faces whose quality is low. Thus techniques such as segmentation or mathematical morphology are complicated yield poor results.

Processing is important as it allows the study of recognition of temporal space patterns in people's faces or individual analysis of the characteristics of the eyes, nose, and mouth. With the applied techniques it is feasible to analyze the behavior of a part of the face without having expensive cameras or search for people on video from the internet.

II. Viola and Jones Method

A. Method

Viola & Jones classifies the image by value with simple characteristics of three types; square entities with sub-boxes called rectangular components [14]. See Fig. 1. These components are shown in Fig. 1, where the grayscale image is scanned by each component to look for positive features with AdaBoost and cascading clarifiers. When a face is detected, a rectangular contour is drawn around the face. The value of the characteristic is the difference between the black and white regions [15]. However, the total number of Haar features is very large, much compared to the number of pixels; AdaBoost is used to select the specific Haar function used and set the threshold value. To ensure rapid classification, the learning process must eliminate most of the available functions and focus on a small set of features [16].

A classification method considers levels of selection. Each level has been trained using the Haar function. The selection separates sub-windows that contain positive objects (images with the desired object) and negative objects (images that do not have the desired object). The Viola-Jones method combined four general rules: Haar Feature, Integral Image, Adaboost learning, and Cascade classifier.

Haar Feature values are obtained from the difference between the numbers of dark areas pixel values minus the number of bright area pixels:

$$\mathrm{F(Haar)} = \sum \mathrm{F}_{\mathrm{white}} - \sum \mathrm{F}_{\mathrm{black}} \qquad (1)$$

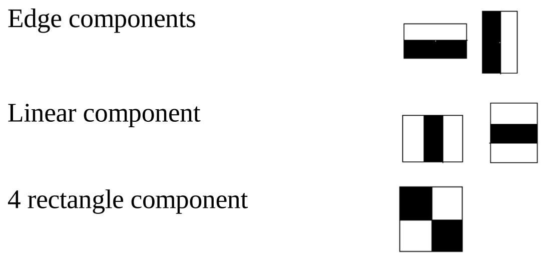


Fig. 1. Haar Rectangular Components.

Where ∑F$_{white}$ is the characteristic value in the brightest area and ∑F$_{Black}$ is the value in the dark area. Haar features are made up of two or three rectangles. The images are scanned and the Haar characteristics in the current stage are searched. The weight and size of each function and the functions themselves are generated using an AdaBoost algorithm [17]. The integral image generated is a technique to quickly calculate the value of the characteristic, and change the value of each pixel to another image representation. See Fig. 2. The integral image in ii(x,y) can be found by equation (2).

$$ii(x,y) = \sum x^{\prime} \le x, y^{\prime} \le y^{i(x^{\prime},y^{\prime})} \quad (2)$$

where ii (x,y) is the integral image at (x,y) and i (x',y') is the pixel value of the original image.

For the overall image in general, small units are added simultaneously, in this case, are pixels. The integral value for each pixel was the sum of all pixels from top to bottom. Starting from the upper left to the lower right, the entire image is the sum with multiple integer operations per pixel. The value of a characteristic is calculated with the value of the integral image at four points. See Fig. 3. If the integral value of the image of point 1 is A, point 2 is A + B, point 3 is A + C, and at point 4 it is A + B + C + D, then the number of pixels in region D is calculated by points 4 + 1 (2 + 3).

The combination of weak classifiers is used to improve the classification. A cascade classifier is a combination of classifiers with a multilevel structure that increases the speed of object detection by focusing only on image areas. See Fig. 4.

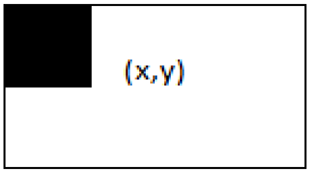


Fig. 2. Integral Image (x, y).

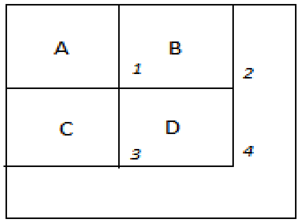


Fig. 3. The Score Count of the Figure.

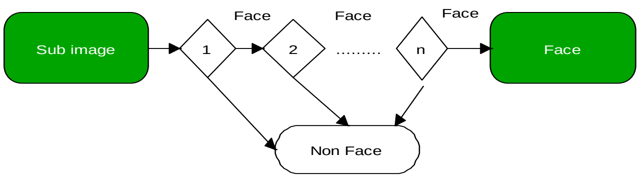


Fig. 4. Cascade Classifier.

A weak classifier is defined by equation (3):

$$h_j(x) = \begin{cases}1, & if\ p_j f_j < p_j\, \theta_j(x) \\ 0, & other\end{cases} \quad (3)$$

Where *ħj (x)* is a weak classification, p_j is even to j, Θ_j is the threshold to j and x is a dimension of the second image. So the strong classifier is:

$$h(x) = \begin{cases}1, \sum_{t=1}^{T} \alpha_t\, h_t(x) \ge \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0, \qquad\qquad\qquad other\end{cases} \; where\ \alpha_t = \log\frac{1}{\beta_t} \quad (4)$$

Facial detection using the Viola & Jones method is presented in Fig. 5. First, the image that can contain a face is read. Second, the Haar feature is processed in the images, which results in the difference of the threshold value of the dark areas and bright areas of the image. If the difference between dark and bright areas is above the threshold value, then there is a face within the image. Face detection is achieved by combining some AdaBoost classifiers as efficient filters that classify the image area. If any of the filters prevents the image from passing, then it has no face. However, when passed through all filters, the image area was classified as a face. The order of the filters in the cascade is determined by the weight given by AdaBoost. The largest weighted filter is placed in the first step, in order to ignore deleting the faceless image area as soon as possible.

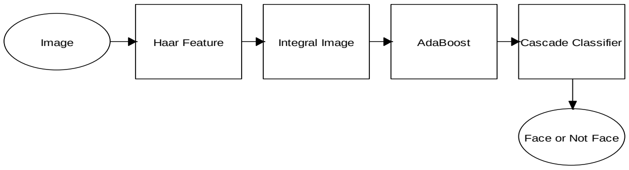


Fig. 5. Face Detection Process with the Viola-Jones Method.

III. Image or Frame Generator System Design

A. Processing

To carry out the processing of the body parts, we must first have the object to be processed, for which a previously stored video file can be opened or a camera that allows the capture of the sequence of frames can also be used. Due to the existence of video files of different dimensions, it is necessary to resize it to observe properly in the user interface. See Fig. 6.

Images can be generated from a video or directly from a webcam, which is useful for comparisons or analysis. See Fig. 7.

The results improve if the face is first located and on that plane, the mouth or eyes are located. Face localization is performed using the Viola & Jones method. It was conceivable to generate images of each of the parts of the face and perform the probabilistic analysis.

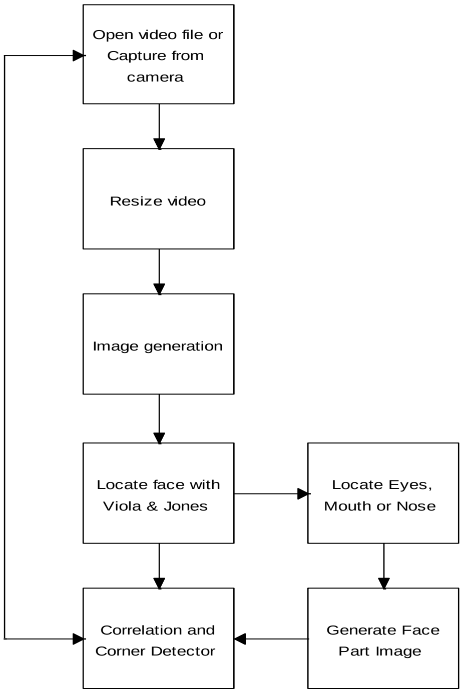

Open video file or Capture from camera
Resize video
Image generation
Locate face with Viola & Jones
Locate Eyes, Mouth or Nose
Correlation and Corner Detector
Generate Face Part Image

Fig. 6. Procedure for Face Analysis.

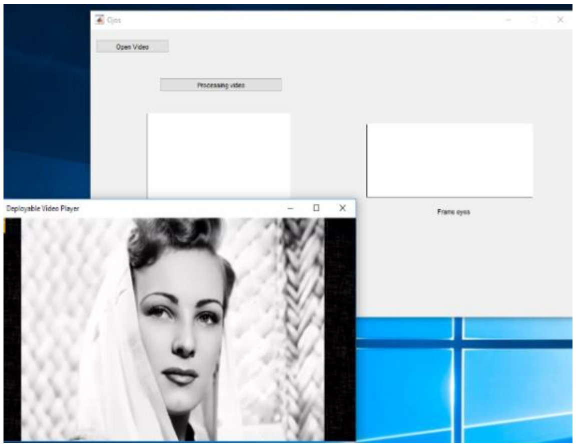


Fig. 7. User Interface “Open Video File”.

Fig. 8 shows the results of locating the eyes of the original video. The high-definition and closeness of the camera to the face make it easy to get good results.

Fig. 9 shows an image of a video with low resolution, full bode and away from the camera; which is very common in internet videos; still, the eyes are located with the method of Viola & Jones; but techniques such as contour detect or segmentation yield poor results. Although, the eyes segmentation is more successful than detector of contour Sobel; however, a histogram threshold is required for each frame or image; it is clear that methods must be rejected and opt for others to detect movement; since the problem can be solved with high resolution cameras as in [18] which increases the cost. Another alternative is the use of specialized hardware based on FPGA, but most people do not have such an option [19]. For this reason, it was decided to solve the problem by means of correlation and algorithm of Harris.

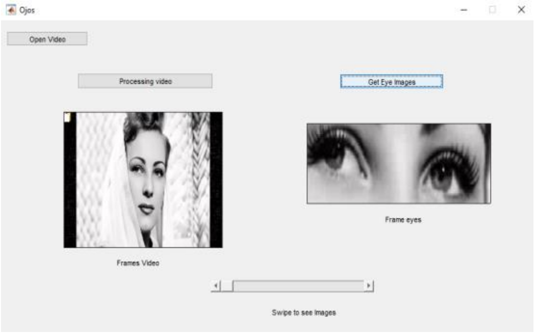
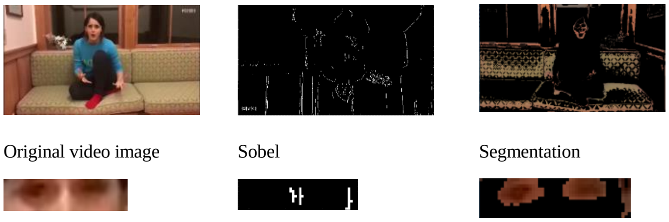


Fig. 8. Obtaining Images.



Original video image Sobel Segmentation

Image of eyes after Viola & Jones Sobel Segmentation

Fig. 9. Images using Sobel Detector and Segmentation.

IV. Analysis of Correlation and Harris using

To complete the project, an analysis was performed by using the correlation between images and the Harris detector The correlation coefficient is calculated by equation (5) [20].

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})\,(B_{mn} - \overline{B})}{\sqrt{\sum_m \sum_n (A_{mn} - \overline{A})^2\,(\sum_m \sum_n (B_{mn} - \overline{B})^2}} \tag{5}$$

Where the means are is equal to $\overline{A} = mean2(A)$, y $\overline{B} = mean2(B)$.

A. Harris-Laplaciano Corner Spot Detector

The corner points in an image are the points that have a significant intensity gradient in the cardinal axes. Harris corners have invariability to rotation and gray level change [21]. Harris's algorithm detects corner points.

It is required to estimate the proper value of the Harris matrix. The Harris matrix is a symmetric matrix similar to a covariance matrix. The main diagonal is composed of the two averages of the square gradients. The elements outside the diagonal are the averages of the cross product of the gradient $\langle G_{xy} \rangle$. Matrix of Harris is:

$$A_{Harris} = \begin{bmatrix} \langle G_x^2 \rangle & \langle G_{xy} \rangle \\ \langle G_{xy} \rangle & \langle G_y^2 \rangle \end{bmatrix} \tag{6}$$

Consider first the measure of the corner response R. The contours of the constant R are shown by thin lines. R is positive in the corner region, negative in the border regions and small in the flat region. Increasing the contrast increases the magnitude of the response. The flat region is specified by T_r, which falls below some selected threshold.

The key simplification of the Harris algorithm is to estimate the proper values of the Harris matrix as a determinant minus the scaled trace or flat squared region.

$$R = \det(A_{Harris}) - k\, T_r^2(A_{Harris)}) \quad (7)$$

Where k is a constant typically with a value of 0.04. R can also be expressed with gradients:

$$R = \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) - k\left(\langle G_x^2\rangle + \langle G_y^2\rangle\right)^2 \quad (8)$$

So that when the response is greater than a predefined threshold, a corner is detected:

$$R > \ k_{thresh}$$

$$\left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) - k\left(\langle G_x^2\rangle + \langle G_y^2\rangle\right)^2 > \ k_{thresh} \quad (9)$$

With this, a pixel in the corner region (positive response) is selected if your response is a local maximum of 8 ways. Similarly, the pixels of the border region are considered edges if their responses are local and negative minimums in the x or y directions, depending on whether the magnitude of the first gradient in the x or y direction, respectively, is greater. This results in thin edges.

By applying low and high thresholds, the hysteresis of the edges can be carried out and this can improve the continuity of the edges. The processing removes stretch marks from the edges and short and isolated edges and joins brief breaks at the edges. This results in continuous fine edges that generally end in the regions of the corners. Edge terminators are then linked to corner pixels that reside within regions, to form a connected edge-vertex graph.

Algorithm of Harris uses equation (7) as a metric, avoiding any division or square root operation. Another way of doing corner detection is to calculate the actual eigenvalues.

The analytical solution for the eigenvalues of a 2x2 matrix can be used in corner detection. When the eigenvalues are positive and large on the same scale, a corner is found. In such a way that:

$$\lambda_1 = \frac{T_r(A)}{2} + \sqrt{\frac{T_r^2(A)}{4}\det(A)} \quad (10)$$

$$\lambda_2 = \frac{T_r(A)}{2} + \sqrt{\frac{T_r^2(A)}{4}}\ \det(A) \quad (11)$$

Substituting the gradients:

$$\lambda_1 = \left(\frac{\langle G_x^2\rangle+\langle G_y^2\rangle}{2}\right) + \sqrt{\left(\frac{\langle G_x^2\rangle+\langle G_y^2\rangle}{2}\right)^2} - \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) \quad (12)$$

$$\lambda_2 = \left(\frac{\langle G_x^2\rangle+\langle G_y^2\rangle}{2}\right) + \sqrt{\left(\frac{\langle G_x^2\rangle+\langle G_y^2\rangle}{2}\right)^2} - \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) \quad (13)$$

V. RESULTS

Tests were conducted with 60 videos taken from the internet, resulting in the detection of movement specifically in faces, eyes, and mouths.

Fig. 10 shows the correlation between consecutive images in case of the complete video. The similarity of the images is observed, due to the value close to 1 of the correlation coefficient; however, abrupt changes can be observed in frame 16 and between frames 53 and 56; caused by camera movements. So we can assume that little sensitive to this event. The statistical data obtained are shown in Table I. This table shows the correlation between consecutive images which proves that it is the same person since the values are positive and close to 1. It is important to check for some cases that it is the same person, since in the videos downloaded from the internet or taken with webcam usually changes the scenario constantly, as well as people recorded or captured on video.

The correlation between the first image and the rest of the images or frames can be made and the result is shown in Fig. 11. The graph shows greater differences, which is corroborated by Table II. This is due to the movement of the person. The range of 0.017 can be seen in the graph of Fig. 11. Although this interval is small, a change can be observed.

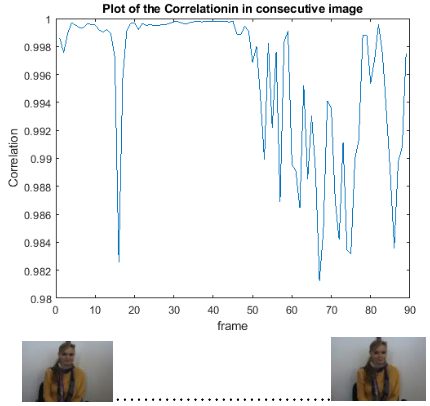


Fig. 10. Consecutive Image Correlation.

TABLE. I. CORRELATION STATISTICS BETWEEN CONSECUTIVE IMAGES

	Frame	Correlación
min	1	0.9812
max	89	0.998
mean	45	0.9958
median	45	0.9989
mode	1	0.9812
std	25.84	0.005309
range	88	0.0186

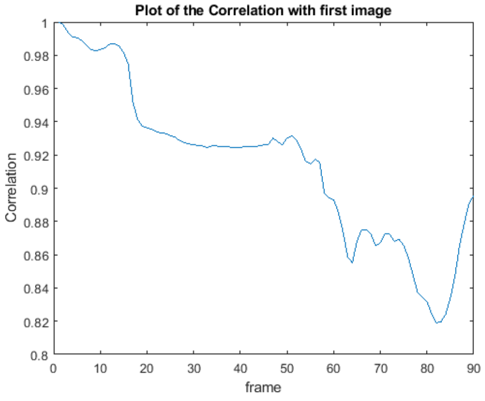


Fig. 11. Correlation between Frame 0 and the Rest of the Images.

TABLE. II. STATISTICS PRODUCED BY CORRELATION

	Frame	Correlation
min	1	0.8189
max	89	0.9986
mean	45	0.9132
median	45	0.8189
mode	1	0.8189
std	25.84	0.04722
range	88	0.01796

The analysis so far considered the total image; now a body part is taken into account. For example, the eyes of the face. For which the Viola & Jones method already described was used and the correlation was applied. Fig. 12 shows the correlation between consecutive images of the eye. It is observed that there are many points that you have no relation some; This is because in the video people open or closes their eyes constantly; for example to blink.

When the images are high resolution or the camera is approached with optical zoom to the human eye, it is possible by morphological operations to obtain the dilation of the iris and thus determine emotional states or some other statistical parameter [22]. However, when you use a webcam or video from somewhere on the Internet, it becomes a more complex process; For this reason, it is better to use corner detectors such as Harris and work with the landmark during the sequence of the event [23][24].

The Harris corner detector was applied to obtain the graph of Fig. 13. Where changes in the number of Landmarks detected are observed; which indicates that there are variations in the different frames processed.

Table III shows the statics of the detected landmarks. There is a minimum of 24 and a maximum of 42 landmarks resulting in a range of 18; which tells us the existence of differences between the number of points detected in these frames; that indicate a change due to the eye movement and that can be corroborated by looking at two images shown in Fig. 13, below the graph.

Without decreasing the number of marks you can also locate the center of the pupils and measure the separation in pixels between the eye. See Fig. 14.

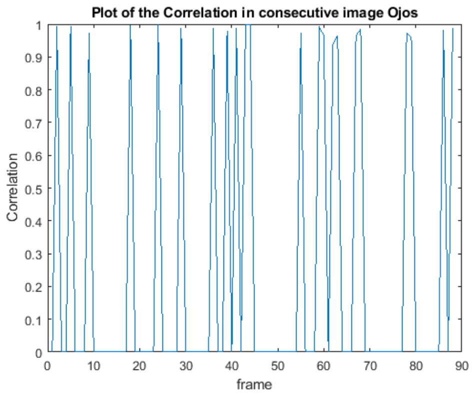


Fig. 12. Consecutive Eye Image Correlation.

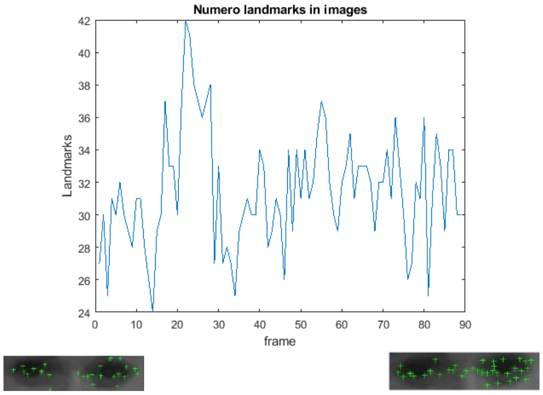


Fig. 13. Landmark Detector by Harris.

TABLE. III. LANDMARKS STATISTICS

	Frame	Landmarks
min	1	24
max	89	42
mean	45	31.51
median	45	31
mode	1	30
std	25.84	3.581
range	88	18

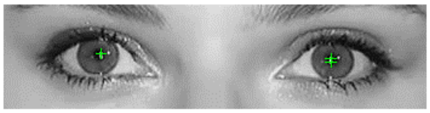


Fig. 14. Separation in Pixels between the Eyes.

VI. CONCLUSION

Movements or changes in the eyes within the image can be identified by correlation in frames or images; however. It is necessary to use methods to detect a change in the images; Changes can be recognized due to eyes movement in correlation graphs. However, the results are better with method of Harris, because the range is greater than that of the correlation.

In the case of the analysis of the eyes, changes were recorded mainly due to the blink of the person; which allows us to conclude that is possible that identify changes in the eyes of the same, even if the video is not high definition or close to the face.

The methods of Viola & Jones and Harris are known and there therefore proven; is for this reason that they are employees. Its use allows motion detection even when you do not have High Definition or hardware at a low cost; since only one web cam is needed, and also is had the advantage of being able to process internet video.

VII. FUTURE WORK

Obtained landmarks by Harris detector, you can use other detectors or methods such as HOG and BRISK, and determine the amount of movement to relate it to affective states or specifically to stress.

ACKNOWLEDGMENT

The authors appreciate the support received by Instituto Politécnico Nacional and Consejo Nacional de Ciencia y Tecnologia.

REFERENCES

[1] D. J. Robertson, R. S. S. Kramer and A. M. Burton, "Face averages enhance user recognition for smartphone security," Plos One, vol. 10, (3), pp. e0119460-e0119460, 2015

[2] T. Chen et al, "Detection of Psychological Stress Using a Hyperspectral Imaging Technique," IEEE Transactions on Affective Computing, vol. 5, (4), pp. 391-405, 2014.

[3] J. Bekios Calfa, J. M. Buenaposada and L. Baumela, "Class–Conditional Probabilistic Principal Component Analysis: Application to Gender Recognition," Class–Conditional Probabilistic Principal Component Analysis: Application to Gender Recognition, 2011.

[4] D. O. Gorodnichy, "Seeing faces in video by computers. Editorial for Special Issue on Face Processing in Video Sequences," Image and Vision Computing, vol. 24, (6), pp. 551-556, 2006.

[5] Zhengrong Yao, Haibo Li, Tracking a detected face with dynamic programming, 1 June 2006, pp.

[6] B. App, C. L. Reed and D. N. McIntosh, "Relative contributions of face and body configurations: Perceiving emotional state and motion intention," Cognition & Emotion, vol. 26, (4), pp. 690-698, 2012.

[7] L. A. Stockdale et al, "Emotionally anesthetized: Media violence induces neural changes during emotional face processing," Social Cognitive and Affective Neuroscience, vol. 10, (10), pp. 1373-1382, 2015.

[8] R. Gonzalez and R. Woods, Digital image processing. New Delhi: Dorling Kindersley, 2014.

[9] K. Rohr and SpringerLink (Online service), Landmark-Based Image Analysis: Using Geometric and Intensity Models. 200121. DOI: 10.1007/978-94-015-9787-6.

[10] M. Codispoti, M. Mazzetti and M. M. Bradley, "Unmasking emotion: Exposure duration and emotional engagement," Psychophysiology, vol. 46, (4), pp. 731-738, 2009.

[11] E. Winarno et al, "Multi-view faces detection using viola-jones method," in 2018, . DOI: 10.1088/1742- 6596/1114/1/012068.

[12] Gonzalo Pajares, Jesús M. de la Cruz, "Visión por computadora, Imágenes Digitales y apliaciones", Alfaomega, México 2002.

[13] E. Cuevas, D. Zaldívar and M. Pérez-Cisneros, Procesamiento digital de imágenes usando MatLAB & Simulink. México, D.F, 2010.

[14] Damanik, Rudolfo Rizki, et al. "An application of viola jones method for face recognition for absence process efficiency." Journal of Physics: Conference Series. Vol. 1007. No. 1. IOP Publishing, 2018.

[15] E. Winarno et al, "Multi-view faces detection using viola-jones method," in 2018, . DOI: 10.1088/1742- 6596/1114/1/012068.

[16] Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume II, Addison-Wesley, 1992, pp. 316-317.

[17] Y. Li, W. Shi and A. Liu, "A Harris Corner Detection Algorithm for Multispectral Images Based on the Correlation," IET Conference Proceedings, 2015.

[18] Smith, M., Maiti, A., Maxwell, A.D., Kist, Colour Histogram Segmentation for Object Tracking in Remote Laboratory Environments A.A. (2020) Lecture Notes in Networks and Systems, 80, pp. 544-563.

[19] Liu, B. Real-Time Video Edge Enhancement IP Core Based on FPGA and Sobel Operator (2020) Advances in Intelligent Systems and Computing, 928, pp. 123-129.

[20] Lewis, J. P., "Fast Normalized Cross-Correlation," Industrial Light & Magic.

[21] D. Yang et al, "A method to detect landmark pairs accurately between intra-patient volumetric medical images," Medical Physics, vol. 44, (11), pp. 5859-5872, 2017.

[22] Herrera, Rodolfo Romero, Francisco Gallegos Funes, and Saul De La O. Torres. "Graphing emotional patterns by dilation of the iris in video sequences." IJACSA Editorial (2011).

[23] Harris, C; and M.J. Stephens. " A combined Corner Edge Detector", Proceedings of the 4th Alvey Vision Conference. Agust 1988, pp. 147-152.

[24] Mikolajczyk, K., Schmid, C. A performance evaluation of local descriptors (2005) IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (10), pp. 1615-1630.