ISSN 2007-9737

Improving the Viola-Jones Algorithm in Thermal Face Recognition of Images

Luis Alberto Hernández Montiel*, Edmundo Bonilla Huerta,
Roberto Morales Caporal, José Crispin Hernández Hernández

Tecnologico Nacional de Mexico,
Instituto Tecnológico de Apizaco,
Mexico

{d23370018,edmundo.bh,roberto.mc,crispin.hh}@apizaco.tecnm.mx.

Abstract. In this paper, a method to improve the performance of Viola-Jones algorithm in finding the relevant parts of a face in thermal images is proposed. First, thermal images are acquired in a semi-controlled environment, and then images containing noise are cleaned. In the next step, the face image is segmented and then enhanced with linguistic modifiers and DIP techniques. Finally, the Viola-Jones algorithm is applied to detect parts of the face such as the mouth, nose, and eyes. The experimental results show that the proposed method is effective, as it achieves a high hit rate.

Keywords. Viola-Jones, thermal image, segmentation, linguistic modifiers, preprocessing.

1 Introduction

Thermal images provide a view of a person from an infrared plane. Working with them can be a somewhat complicated task, as these images have specific hot and cold shades and are sensitive to different environmental conditions. Finding an interest region in the thermal image can be very helpful in diagnosing a disease or identifying a specific emotion in a face that is not seen in a visible plane image. Finding these interest regions within the thermal image is a difficult task because many segmentation algorithms can get confused and select a region that is not the correct one or do not select anything. Nowadays, several methods or improvements of segmentation algorithms have been proposed to successfully detect the interest region. To solve this problem, this paper proposes a method based on DIP techniques and linguistic modifiers to enhance the thermal image so that the Viola-Jones algorithm can detect the interest regions. First, a segmentation method is developed to find a face in 250 thermal images of students aged 18 to 21 years old. Then, the image is enhanced with linguistic modifiers and DIP techniques. Finally, the Viola-Jones algorithm is used to detect and segment the regions of interest in the image. This method will be used to segment the eyes, nose, and mouth in 250 thermal facial images.

2 Related Work

The thermal facial image study to identify, recognize and classify emotions in facial features, is an opportunity area that must be evaluated and studied from different approaches. Thermal facial imaging is limited in terms of subject-to-camera distance, temperature, and light variations.

Therefore, it represents a computational challenge to propose algorithms or improving methods that can classify basic facial emotions in thermal images due to low-contrast and intensity limitations. To know the conditions for thermal infrared facial images, several studies have been reported in the literature.

In a study proposed by [1] the importance of the process of collecting facial data from a thermal camera is described. Some aspects related to the image capture environment are considered, such as the ambient temperature and the temperature

of an individual, air flow, humidity, the focal length of the camera, and the subject's position during the tacking capture of the thermal image of each subject.

Similar work is reported by [5] where students' emotional states are considered during the learning process by using non-invasive technology. This study recollected articles from different digital databases from 2010 to 2022 to analyze the advantages and disadvantages of effective computing in education using thermography and deep learning techniques. In the last decade, different approaches have been proposed to tackle the problem of low contrast and lower image resolution in thermal images.

In [12] a framework composed of two stages was proposed: The first stage consists of an extraction feature of blood vessels from thermal facial images, based on four steps: a) face segmentation, b) unwanted noise removal, c) morphological operators, and d) postprocessing to obtain thermal signatures. In the second stage, a similar thermal signature metric signature is proposed for face recognition.

Heet et al. [7] proposed an original Deep Boltzmann Machine (DBM) to learn thermal features for facial expression recognition. The regions to extract discriminating emotions such as disgust, fear, and happiness ON the forehead, mouth, and cheek. Other regions of the face are considered irrelevant in this study. In the first step, the main regions are located and a thermal images normalization technique is applied. In the second step, DBM is proposed for facial recognition expressions.

In [11] feature extraction and several techniques such as deep learning, random forest, and ensemble classifiers for thermal facial images were reported. This crucial step is the feature extraction process to obtain several blocks of the facial image and generate a new feature image. In the classification prediction of this major architecture, vote techniques are applied.

A gentle review is presented in [9] for thermography active and passive-based machine vision with a deep learning approach. In case of thermal facial recognition systems, an inception V3 neural network is reported for fast face tracking. Face alignment is also mentioned by using the landmark technique or a set of landmarks.

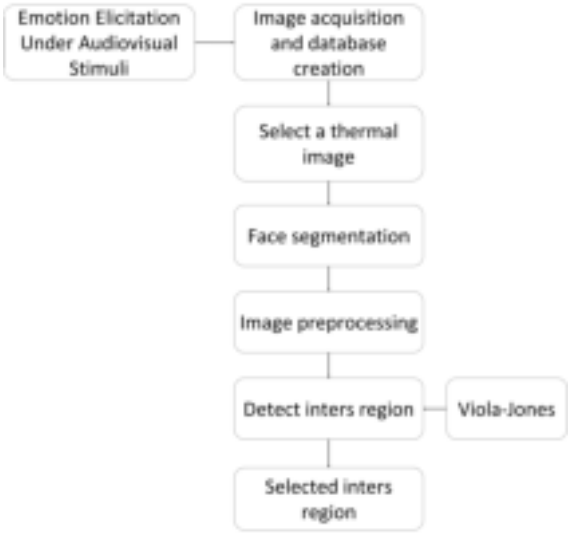

Fig. 1. Proposed model general scheme

3 Materials and Methods

Working with images that are not in the visible plane becomes a complex problem because a thermal or infrared image generates images from the heat emanating from the object. Thus, if the object is exposed to a higher heat source than its own, the result of the image may be different from expected and may contain noise.

A thermal imaging camera operates in the infrared plane (0.78 µm to 1.7 µm) and a digital camera operates in the visible plane (0.38 µm to 0.78 µm). This means that the images captured by the thermal imaging camera are more sensitive to small changes in the environment (light, air, heat).

The thermal image processing problem has been little addressed because conventional object detection (DIP) algorithms tend to make errors because they have a large search area with less accurate information. To address this problem, this section presents the proposed model that combines DIP techniques with linguistic modifiers to search for face parts in thermal images. Figure 1 shows the general flow of the algorithm.

3.1 Triggering Emotions by Audiovisual Stimuli

Throughout the day, people may experience different emotions in response to situations they find themselves in day after day. Basic emotions like disgust, anger, fear, sadness, amusement, tenderness, and neutrality are the most expressed, as well as admiration, adoration, fear, amazement, discomfort, and boredom, among others [3]. One of the methods to generate these emotions is video stimulation. When a viewer watches a video that contains a horror or fun scene, this can cause the person to manifest an emotion through what they see. In our case, we used videos from the LATEMO-E database [14]. This DB contains 28 video clips or movie fragments with an average duration of 152 seconds, covering topics such as fear, joy, or sadness, among others. The selection criteria for these videos were the following: (1) the most suggested movies, (2) availability on YouTube or in video clubs in Latin American language, and (3) the coherence of the action in the scene with the addressed topic. These video clips should evoke a kind of basic mood in the subjects to be photographed.

3.2 Thermal Image Acquisition and Database Creation

Thermal imaging uses object temperature to create an image in a non-visible plane [16]. For this study, thermal images were acquired under the following conditions: The images were captured using a Fluke IR -Fusion® Technology camera [8]. This is a camera that simultaneously captures a digital image in visible light and an infrared image. For this study, 480 images were taken, of which only 250 were used. 25 subjects participated, 13 mens and 12 females between the ages of 18 and 21 years old. The images were taken in a semi-controlled environment, within 1.5 meters from the camera and the subject, and at a height of 1.2 meters from the floor to the camera. The subject sat in front of a screen on which videos were transmitted to evoke facial emotions. We used Bose headphones to isolate the sounds and to be able to express themselves freely and to obtain different recordings of the same person. Figure 2 shows the environment in which the images were taken.

Fig. 2. Working scenario for taking photographs

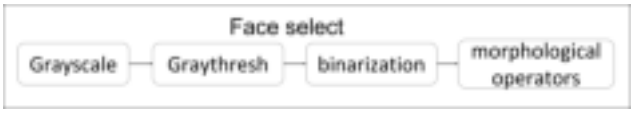

Fig. 3. face segmentation process

3.3 Thermal Image Selection

In this phase, the thermal images with which the algorithm will work were selected. Several images of each participant were taken to create a database of different reactions on the person's face (laughter, seriousness, anger). A total of 480 images were taken, not all of which were useful for the study. A simple cleanup eliminated the images exposed to a very strong light or heat source and discarded the images where the participant's face was not visible or only partially visible. The result is 250 effective images that can be analyzed.

3.4 Face selection

In this phase, the person's face is searched in the thermal image. The idea is to locate the object of interest in the image and segment it by eliminating the rest of the image, which is considered noise. For this purpose, a method is proposed that uses different DIP techniques, which are described below. Figure 3 shows the facial segmentation process.

3.4.1 Grayscale

In this phase, the thermal image is converted to a grayscale image with more neutral color tones to process it more efficiently. The thermal image is converted to a grayscale image as follows. The color equivalency within the image is calculated as a measure for each pixel. For this purpose, the luminance method [19] is used, which is a mathematical expression that represents the intensity with which the human eye perceives the spectrum frequencies near the RGB color scale and the illumination of these colors (RGB) within the image. Then, each image pixel is decomposed into its three colors (RGB), and by applying equation 1, the pixel is given its equivalent in grayscale [19]:

$$Y = R * 0.3 + G * 0.59 + B * 0.11, \tag{1}$$

where: Y is the grayscale output image, R is for Red, G is for Green, and B is for Blue. The values 0.3, 0.59, and 0.11 are standard values established by the *International Telecommunication Union - Radiocommunications* [19].

3.4.2 Otsu Threshold

In this phase, the Otsu method [4] is used to find the optimal threshold. This method uses variance as a measure of dispersion so that the gray levels within each class are as small as possible, but at the same time as large as possible between different classes. To find a threshold (t) and double a bimodal image, the entire image histogram is traversed, and the minimum variances between black and white are calculated. The threshold (t) is the sum of the variances of the two classes according to the equation 2 [4]:

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t), \tag{2}$$

where: $\sigma_w^2(t)$ is the expected threshold. $q_1(t)\sigma_1^2(t)$ is the variance of class 1 and $q_2(t)\sigma_2^2(t)$ is the variance of class 2. This method determines an automatic threshold that divides the whole image histogram and manages to distinguish the pixels distributed in both classes, which helps the binarizing process.

3.4.3 Binarization

Image binarization is the conversion of grayscale images into black-and-white images. It is a basic digital image processing technique that separates the background from the objects of interest by using a threshold that divides the histogram into two classes (black and white) as follows. If the pixel is greater than or equal to the selected threshold, it belongs to the white side; if it is less than the threshold, it belongs to the dark side. Equation 3 shows how binarization works [20]:

$$(I_{bw}) = t(I_g)Where \begin{cases} 0, & \text{if } I_{(i)} \geq t \\ 255, & \text{if } I_{(i)} < t, \end{cases} \tag{3}$$

where I_{bw} is the binarized image. t is the Otsu threshold. (I_g) is the grayscale image. The threshold is used as a discrimination method to convert an image with continuous pixels (grayscale) to an image with black-and-white or discreet pixels. One of the problems with binarizing the image is that it can produce small pixels with a different color than the region (noise). To remove this noise, we applied morphological operators that effectively cleaned up the image. The morphological operators are described below.

3.4.4 Morphological Operators

Morphological operators are a wide range of mathematical operations that apply a structural element to an input image and produce an output image of the same size [23]. The basic operators are dilation and erosion. Dilation adds pixels to the object boundaries of an image, while erosion removes pixels from the object boundaries. The number of pixels added or removed depends on the size and shape of the structural element used to process the image. In our case, we use the morphological aperture operator, which consists of an erosion operator followed by a dilation operator, using a disk structural element for both operations. The aperture operation is defined by [23]:

$$X \circ Y = (X \ominus Y) \oplus Y, \tag{4}$$

where X is the binary image, Y is the structural element, $\circ$ means the erosion operator, and $\oplus$ means the dilation operator. With the

morphological operators, it is possible to improve the image by removing small perturbations that may occur during binarization. This new image can be used as a marker to segment the regions of interest within the original image. First, the binarized image is pasted over the grayscale or color image. Then we go through the binary image and evaluate each bit. If the bit is active (1), we select the pixel located at the position associated with the active bit; otherwise, we set 0.

3.5 Image Preprocessing

Preprocessing or data cleaning is the process of removing data or noise that can interfere with data analysis and lead to erroneous results. In our case, image preprocessing is performed by combining fuzzy linguistic modifiers with histogram adjustment to improve the image. The goal of this phase is to eliminate the noise, improve lighting, and highlight the part of the image of interest. This phase is an improvement of the image obtained in the previous phase. The next step is to apply a linguistic modifier to the new image to highlight the interesting parts, and the last step is to adjust the histogram of the image to balance the contrast. The whole process is shown in Figure 4.

Fig. 4. Imagen preprocessing process

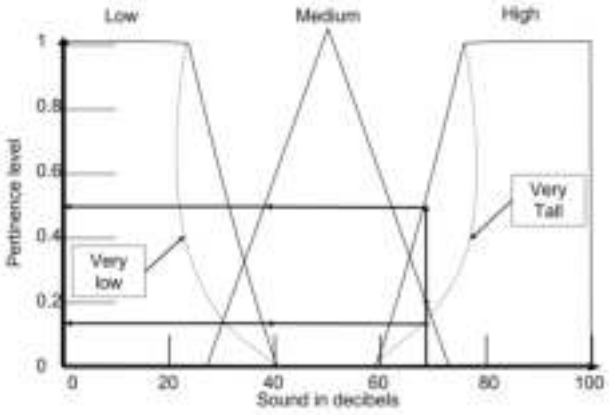

Fig. 5. Application of the linguistic modifier ”very” to the low and high sets

3.5.1 Linguistic Modifiers

After segmenting the face within the image, the next step is to enhance the image by highlighting its most distinctive features (eyes, nose contour, lips). For this purpose, linguistic modifiers are implemented. These modifiers are used to change the shape of fuzzy sets. They can be associated with adverbs such as ’very’, ’somewhat’, and ’a little’ [15].]. For this purpose, a set of simple equations is used, which can improve or decrease the membership degree of a fuzzy set. An example of the use of linguistic modifiers can be found in Figure 5.

As can be seen in Figure 5, the membership level of the sound changes when the linguistic modifier is applied since the sound is a high-setting element with a 0.5 relevance level, but it also belongs to very high settings with a 0.18 relevance level. There are different linguistic modifiers that can expand (weak modifier) or restrict (strong modifier) the fuzzy set. The most widely used modifiers are concentration, dilation, and intensification, but they are not the only ones; there are others that are modifications of classical linguistic operators. These include the modifiers High, Quiet, and Plus, which modify the concentration operator to further emphasize the result. Other operators combine the classical modifiers with logical connectors to enhance the result. These include Rather, Slightly and Pretty, which use logical operators such as negation to achieve a better result than that of the classical linguistic modifier [15]. In our case, we used the linguistic modifiers Concentration, Dilation, and Intensification, and some of their variations were used, combining them in different ways to enhance the image and highlight the face in the images efficiently.

3.5.2 Image Adjustment

In the last phase of preprocessing, the goal is to adjust image contrast to obtain a more balanced image between light and dark. This is done by manipulating the image histogram. A histogram shows the tonal values taken by each pixel within the grayscale image. By viewing the image histogram, you can also detect noise or interference in the image. A very bright image means that most pixels are on the white side, while a very dark image indicates that the pixels are on the shadow side [22]. To solve this problem, there are several methods to adjust the image contrast. The idea is to reallocate the intensity values so that all pixels are distributed within the histogram. That is, if the image brightness is too high, all pixels in the image become brighter, and vice versa, if the brightness is too low, all pixels in the image become darker. Thus, the intensity of the image area is balanced and the processing algorithms achieve better results. There are several linear and non-linear methods to perform this process. In this case, the following formula is applied to perform the adjustment of the image obtained in the previous process [22]:

$$g = \left(\frac{(f-a)}{(b-a}\right)^{\gamma} * ((c-d)+d), \tag{5}$$

where g is the output image, f is the input image, $a = low\_in$, and $b = high\_in$ are the minimum and maximum pixels within f, gamma (γ) is a value that indicates where the output should be loaded more, white, or dark if gamma has no value, the default value of 1 is assumed. $d = low\_out$ and $c = high\_out$ are the set values of 0 and 1. Applying Equation 5, the intensity of the image is balanced and ready for the next step. At this point, the preprocessing is complete and a new image with better properties is generated, which is evaluated to find the different parts of the face.

3.6 Interest Region Recognition

In this phase, the regions of interest are searched in the face of the subject. For this purpose, the Viola-Jones algorithm is used [18]. The original image goes through segmentation and facial preprocessing.

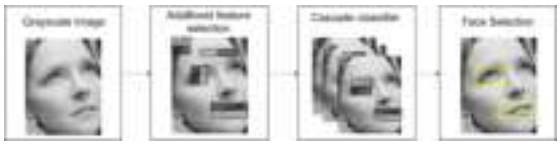

Fig. 6. Process of the Viola-Jones algorithm

An enhanced image is obtained in which only the person's face is visible. This new image is the input image for the Viola-Jones algorithm. The Viola-Jones algorithm was proposed by Paul Viola of Mitsubishi Electric Research Labs and Michael Jones of Compaq CRL in July 2001 [18]. This algorithm processes the information within a grayscale image as follows. First, it uses an image called an integral image, which is a copy of the input image. To search for an area of interest in the image, the algorithm plots different subregions with different sizes over the integral image. Within these sub-regions are the set of pixels that would be used for the evaluation. The next step is to create an adaboost classifier.

To speed up the classification, the algorithm combines weak classifiers that exclude most of the erroneous pixels and focus on a small group of relevant pixels. In this way, a boost process is generated that serves as a feature selection method [6]. Then, a series of more complex classifiers are used in a cascade. The pixels are evaluated at each classifier stage by determining if the subregion belongs to the interest area. If the result is negative, that subarea is discarded and another subarea is evaluated. If the result is positive, the subregion is sent to the next stage to be evaluated, and so on until the last stage of the classifier is reached. If the result of the last stage is positive, it means that the algorithm has detected the interest region. This increases the classification speed, as the classifiers focus only on the subareas of interest in the image, eliminating the likelihood of selecting a false positive that does not match a face [6]. Figure 6 shows the Viola-Jones algorithm flags a single eye and the mouth as interest regions within the image.

Once the regions are identified, the next step is to segment the selected regions.

3.7 Interest Region Selection

Once the region is detected, the final step is to segment that region. Segmentation consists of dividing the image into its parts of interest. In our case, we are interested in the person's features (eyes, nose, or mouth). For this purpose, we first obtain the coordinates of the region detected by the Viola-Jones algorithm in the previous step. We cropped the image using these coordinates and obtained a new image. The next step is to find a threshold for image binarization. The threshold is found by applying Equation 2, the Otsu method. Equation 3 is then applied to remove the background and obtain only the binarized interest region. This image is used as a mask to highlight parts of the face in the original image. At this point, one cycle has been completed, and the algorithm is ready to receive a new image.

4 Experiments and Results Analysis

In this section, we describe the experiments performed with the proposed algorithm. The experimental protocol was performed on an HP laptop with AMD Ryzen 7 processor and 8GB RAM. The algorithm was implemented in Matlab version R2021a. The best results for each phase are described below.

4.1 Images were Taken with the Fluke IR-Fusion® Thermal Imaging Camera

Thermal images, or in the non-visible plane, are created from infrared radiation emission given off by an object as a function of its temperature, i.e., the higher the object's temperature, the stronger the infrared radiation emission. The cameras that record this radiation print an image of the object in the colors of the temperature it emits. In our case, a Fluke IR-Fusion® camera was used and the images obtained can be seen in Figure 7.

Figure 7 shows only 9 examples of the 250 images we worked with. All images have a standard size of 480*640 pixels and are in JPG format. The images were taken of 25 students (as mentioned in the previous section) during the staggered return to school. The colors show the different temperature changes in the student's face and body. Figure 8 shows the range of colors captured in the image.

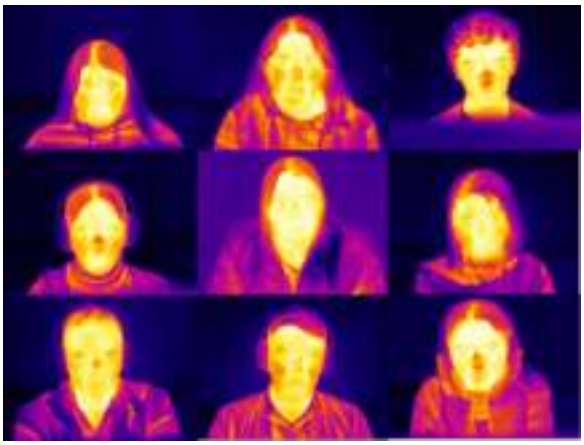

Fig. 7. Image samples obtained by the Fluke IR-Fusion® camera

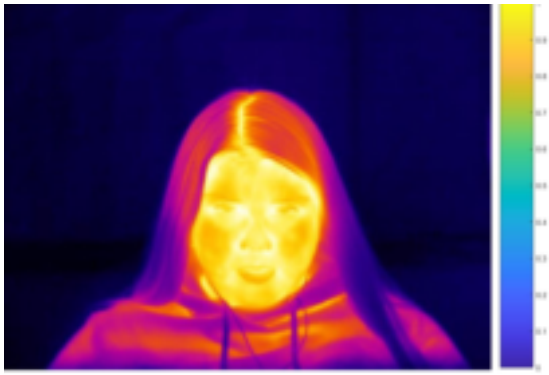

Fig. 8. The color range of the thermal image

The colors shown in Figure 8 reflect the various temperature changes in the student's face and body. The shades from blue to dark represent all the cold areas of the image, such as the hair and the background. The shades from orange to yellow indicate the warm areas of the image, such as the face and its features, such as the eyes, nose, mouth, and part of the head.

4.2 Facial Segmentation Results

In this phase, the person's face is determined by combining different DIP techniques. First, the image is converted to grayscale, and then a threshold is determined to serve as a reference for image binarization. In the last step, morphological operators are used to enhance and denoise the image. The results of each phase are described below. A grayscale image is the representation of an RGB color image in shades of gray. This conversion is done with the intention of reducing the image size so that the segmentation algorithms can work better on it. To achieve this conversion, there are several mathematical expressions that take the color pixel and look for its equivalent in grayscale. In our case, the luminance method [19] is used and the results are shown in Figure 10.

As can be seen in Figure 10, pixels that are in the RGB model color region get their equivalent in grayscale using the luminance method. Thus, an image with three dimensions (one for red, one for green, and one for blue) becomes a two-dimensional image contains the gray luminance levels present in each pixel. Figure 9 shows the gray areas contained in each image.

As can be seen in Figure 9, the gray palette in the image ranges from 0 to 255. The dark colors range from 0 to 50 representing the background, hair, and some of the clothing in the image, while the lighter colors range from 200 to 255, representing face contours, eyes, and nose. The gray shades between 100 and 200 correspond to the forehead, cheekbones, nose, mouth, and eyes in the image.

Once we have the images in grayscale, the next step is to determine a threshold to divide the image into two classes: black and white. To do this, we use the Otsu threshold, which traverses the entire image histogram, to calculate the variances between black and white, sum them up, and use them as a discriminating value. The results of this step can be seen in Figure 11.

Figure 11 shows three grayscale images in the upper part. The lower part shows the histograms of each image. The threshold value determined by Otsu is marked with a red line that divides the histogram into two classes. This value is used as a distinguishing value for image binarization. The threshold values of the three images are shown in 1 as follows. The first column shows, from left to right, the image number corresponding to Figure 11. The second column shows the threshold value obtained using the Otsu method.

Table 1. The threshold value was determined with the Otsu method for Figure 11

Image number	Threshold
Image 1	Pixel 118
Image 2	Pixel 121
Image 3	Pixel 100

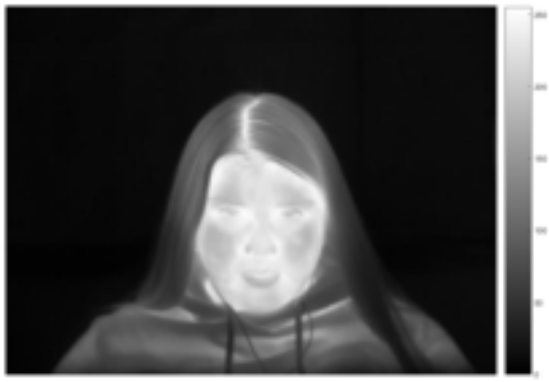

Fig. 9. Gray ranges within the image

Table 1 shows how the thresholds change depending on the image being evaluated. Each threshold is different for each processed image, resulting in 250 different thresholds. This is done because each image has its own gray levels, and if a default threshold is used, interesting information could be lost when binarizing the image. The next step is to binarize the image. The threshold of Table 1 is used as the discriminating value when applying Equation 3. In this way, we can convert the pixels to a white or black color and obtain an image consisting only of a matrix of 0 and 1. Figure 12 shows the results of some binarized images.

Figure 12 shows that when the image is binarized, only the region of interest remains, although there are still parts marked as interest regions that are

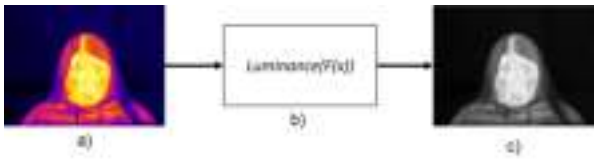

Fig. 10. Image conversion from color to grayscale. a) is the color image, b) is the luminance method, and c) is the image converted to grayscale

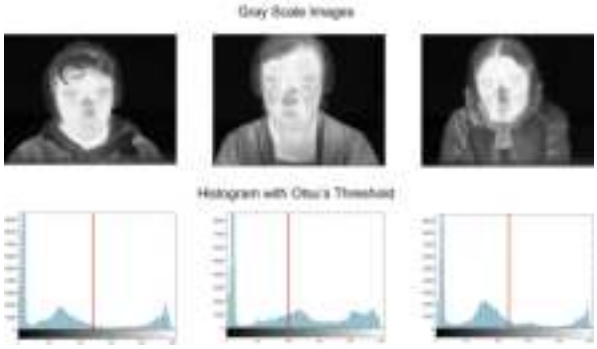

Fig. 11. Threshold within the image histogram

Fig. 12. Binarized images

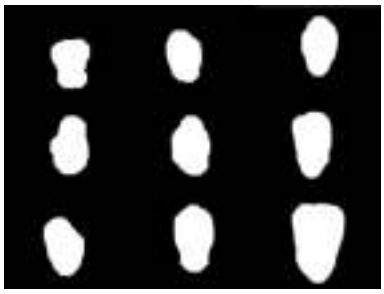

Fig. 13. Morphological application results

considered noise. To remove this noise, we used the morphological aperture operator (Equation 4), which first performs erosion and then dilation. This allows us to eliminate the uninteresting parts of the image and keep the largest area only. The phase results can be seen in Figure 13.

The area shown in Figure 13 is the outline of the person's face. This binary image is used as a mask to obtain the person's face in the grayscale image. For this purpose, the grayscale image pixels are selected, using the bits that are active (1) in the binary image as a reference. The result of this process is shown in Figure 14.

If we only have the person's face, we can be more precise in finding features such as eyes, nose, and mouth. The total faces found using the proposed segmentation method are shown in Table 2 as follows. Column 2 shows the results obtained with the algorithm. Column 3 shows the percentage of hits obtained by the algorithm.

Of the 250 images used to train the proposed algorithm, a total of 246 complete faces were segmented, resulting in an accuracy rate of 98.4%. The algorithm achieved an error of 1.6%, which means that it confuses white pixels with dark pixels and vice versa. Figure 15 shows the errors that the algorithm had during segmentation. The algorithm detected only part of the face during segmentation, so it is not possible to work with it because interesting parts like the nose or mouth are missing. These images were discarded for the next process, and only the 246 faces that were correctly segmented were worked with.

4.3 Preprocessing

In the preprocessing phase, the image is treated to remove noise or light interference (strong or weak) that occurred when the image was captured and to highlight important features or regions within the image to make them easier to locate in the next phase. In our case, we used preprocessing that combines various DIP techniques with linguistic modifiers. The results for each stage are shown below. In this stage, we worked with the images obtained in the previous stage and analyzed only the person's face. First, fuzzy linguistic modifiers were used to enhance the image obtained in the

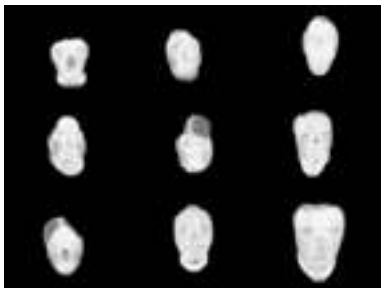

Fig. 14. Segmentation of the face in a grayscale image

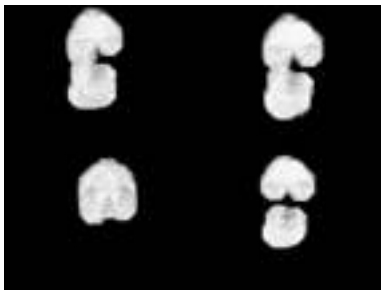

Fig. 15. Erros in te face segmentation

previous stage. The phase goal was to enhance the features of the person in the image. There are different linguistic modifiers, and each one manipulates the image in a different way. In our case, we used a combination of basic modifiers to improve the features of the image. The best combinations are shown in Table 3 as follows. Column one shows the number of the combination and column two shows the name of the modifiers used. Figure 16 shows the images processed with these operators.

We combined these modifiers because processing the image with one modifier improved one region and other modifiers highlighted another. By combining these modifiers, we were able to highlight regions such as eyes, nose, and mouth in the image, as shown in Figure 16.After applying the linguistic modifier, the next step was image adjustment. In this step, the image histogram was adjusted using Equation 5. By adjusting the image, the distribution of whites and darks within the image was balanced and fully distributed across its histogram. The results of this phase can be seen in Figure 17.

Fig. 16. Image enhancement using the modifiers in Table 3

Table 2. The total number of segmented faces

	Total	Percentage
Color images	250	100%
Segmented Faces	246	98.4%
Errors	4	1.6%

As can be seen in Figure 17, when adjusting the contrast of the input image (left image), the grayscale intensity values are assigned to the input image using a default lower and upper saturation of 1%. This process increases the contrast of the output image (right image). This contrast adjustment is performed for all images from the previous step. With this last step, preprocessing is complete, and the image is ready to locate the parts of interest. The best results are described below.

4.4 Interest Region Detection

In this phase, interest regions were searched for within the enhanced regions. This process was performed using the Viola-Jones algorithm. The algorithm goes through each part of the face looking for the most important regions (nose, mouth, eyes). Once it finds them, it uses a cascade classifier to make sure it is the interest region, selects it, and tags it. Figure 18 shows the results for the nose region.

Figure 18 shows the nasal region selected by the Viola-Jones algorithm for each of the combinations listed in table 3 plus the histogram adjustment. This produces five results for each interest region. Table 4 shows the results obtained for the nasal

Table 3. Linguistic modifier combinations

Number	Combination
1	slightly + plus
2	Rather + Minus
3	Pretty + FAIRLY
4	Pretty + Dilatation
5	slightly + Intensification

Table 4. Results for the nose region

Combination	Nose	Percentage
slightly + plus	234	95.1%
Rather + Minus	221	89.8%
Pretty + FAIRLY	190	77.2%
Pretty + Dilatation	207	84.1%
slightly + Intensification	179	72.8%

region. The first column shows the name of each combination, column 2 shows the number of images processed correctly, and column 3 shows the algorithm success percentage.

Table 4 shows the proposed method effectiveness percentage for the nasal region. The combination of morphological operators, reaching a percentage of 95.1%, detects 234 out of 246 possible noses and produces only 4.9% errors, i.e., 12 noses that were not detected in the desired region. Another region that was searched for was the eyes. The results for this region are shown in Figure 19.

As shown in Figure 19, the Viola-Jones algorithm effectively selected and labeled the eye region in each of the images resulting from the previous step. Table 5 shows the results obtained for the eye region.

Table 5 shows that the best percentage was slightly + intensification with 89.4% of hits, finding 220 pairs of eyes among the 246 possibilities. The error was 10.6%, i.e., a total of 26 images in which the eye region was not detected. The final algorithm test was to find the mouth region. For the eyes and nose, the algorithm was applied to the five results of the preprocessing phase.

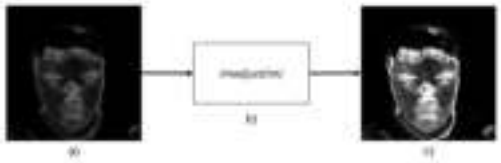

Fig. 17. Enhancement of image regions with contrast adjustment. a) image processed with linguistic modifiers, b) histogram adjustment, c) image with adjusted contrast

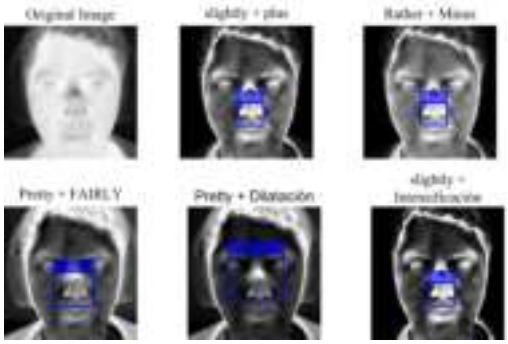

Fig. 18. Selected nasal region

Figure 20, shows the result for the mouth region.

As you can see in Figure 20, the mouth region was successfully identified. It is worth noting that in this region, unlike the eyes and nose, the Viola-Jones algorithm confused some other parts of the mouth. The results for this region are shown in Table 6.

As shown in Table 6, the percentages of hits are somewhat low compared to the other regions. The best result was obtained by slightly + intensification with 82.1%, corresponding to 202 mouths found and an error of 17.9%, i.e., 44 images were confused with other regions. Table 7 shows the overall percentage of the proposed method for each selected region. Column 1 shows the proposed method, column 2 shows the results for the nose region, column 3 shows the results for the ocular region, and column 4 shows the results for the mouth region. Each column shows two results, ac for the percentage hits and er for the error of the algorithm.

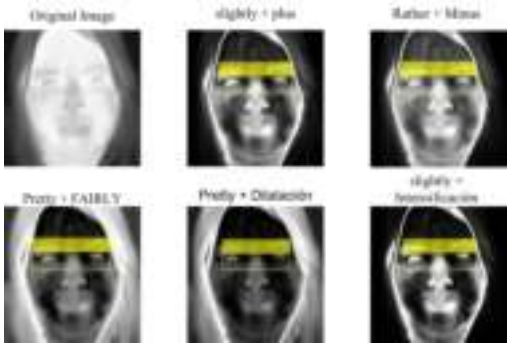

Fig. 19. Selected eye region

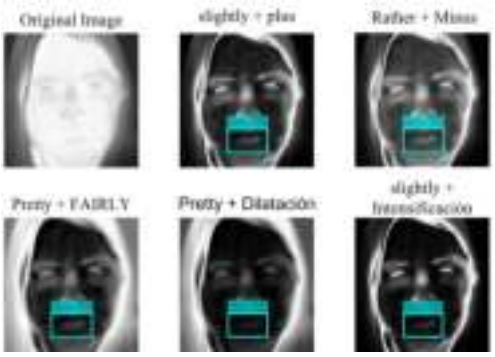

Fig. 20. Selected mouth region

Table 5. Results for the eye region

Combination	Eyes	Percentage
slightly + plus	205	83.4%
Rather + Minus	200	81.3%
Pretty + FAIRLY	196	79.8%
Pretty + Dilatation	165	67.1%
slightly + Intensification	220	89.4%

Table 6. Results for the mouth region

Combination	Mouth	Percentage
slightly + plus	198	80.4%
Rather + Minus	156	63.4%
Pretty + FAIRLY	85	34.5%
Pretty + Dilatation	143	58.1%
slightly + Intensification	202	82.1%

The results in Table 7 show that the best performance was obtained in the nose region with an overall average of 83.8%, corresponding to a total of 206 noses detected on average. For the eye region, the algorithm achieved an overall average performance of 80.2% and selected an average of 197 pairs of eyes. For the mouth region, the algorithm achieved an overall average of 63.7%, corresponding to an average of 157 mouths. To obtain this average, the results obtained for each region were added and divided by the number of tests. So far, the proposed method has found interesting regions. The next step is to binarize and extract them from the original image.

4.5 Selecting the Interest Regions

After the Viola-Jones algorithm has detected (marked) the interest regions in the images, the next step is to select and segment these regions. The next step is to crop the image to preserve only the selected region. To do this, the Matlab imcrop function is used as follows. The Viola-Jones algorithm returns 4 data to mark the region found on the image, this data is:

$$[x, y, h, w] = vj(img), \tag{6}$$

where x is the coordinate on the x-axis, y is the coordinate on the y-axis, h is the height, and w is the width. These four values are sent to the imcrop function to crop only the part of interest, as shown in Figure 21. Once the image is cropped, the next step is to determine the threshold by applying Equation 2 of the Otsu method. In this method, all pixels are evaluated in gray levels to select a numerical threshold (t) that helps to binarize the image. After setting the threshold (t), the next step is to binarize the image. We use the threshold (t) to assign pixels to a white (0) or black (1) value. Applying Equation 3 to the image, we can segment the region of interest. Figure 22 shows the process of binarizing the image.

As can be seen in Figure 22, the region is correctly segmented, eliminating the entire image and leaving only the nose. This process also applies to the eyes and mouth regions. Figure 23

Table 7. Overall average of the proposed approach for each region

Region					
Nose		**Eyes**		**Eyes**	
ac	er	ac	er	ac	er
83.8%	16.2%	80.2%	19.8%	63.7%	36.3%

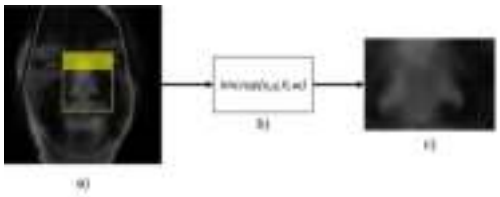

Fig. 21. Cropped nose region. a) Selected region. b) Method of cropping the image. c) Cropped region

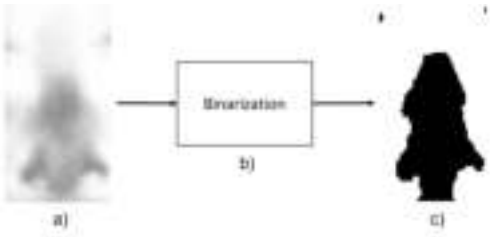

Fig. 22. Binarized image. a) cropped area. b) binarization method. c) Binarized image

shows the result of this process. As can be seen in

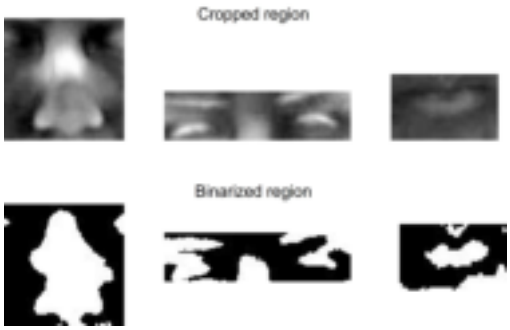

Fig. 23. Cropped and binarized image

Figure 23, each region has been segmented and binarized for all images from the previous phase. As you can see, some images have a slight noise. We will try to improve the binarization so that only the interesting regions remain.

4.6 Comparative Study

Two comparative studies were performed. In the first, two tests were performed using the Viola-Jones algorithm without the segmentation and preprocessing phases. In the first test, the color image was analyzed to see if the algorithm was able to locate the interest regions within the image. In the second test, the image was converted to grayscale and analyzed using the Viola-Jones algorithm. The results of these two tests are shown in Table 8. The first column contains the name of the region searched, the second column shows the results of the Viola-Jones algorithm when working with color images, and the third column shows the results of the algorithm when working with grayscale images.

As shown in Table 8, the results in both tests are low, indicating that the Viola-Jones algorithm does not work correctly with thermal images and does not correctly find the interest regions because it skips them or assumes another region as the interest region. Table 9 shows the comparison of the Viola-Jones algorithm results with the proposed method.

To compare the Viola-Jones algorithm results with the proposed method, it can be concluded that the proposed method improves by finding the interest regions within the image more accurately. In the second comparison study, the performance of the proposed method was evaluated with different methods described in the literature. Table 10 shows the comparison study. The table is divided as follows: The first column shows the method it was compared with and its reference. Columns 2, 3, 4, and 5 show the performance found for each interest region.

5 Conclusions

In this paper, a method to process thermal images and find the parts of the face such as the mouth, nose, and eyes was proposed. First, the algorithm takes an image of the student while Stimulating him or her with different videos to generate facial

Table 8. Viola-Jones algorithm results with the raw images

Region	Images	
	Color	Grayscale
Faces	146(58.4%)	**185(74%)**
Nose	66(26.4%)	**69(27.6%)**
Eyes	0(0%)	**1(0.004%)**
Mouth	**70(28%)**	41(16.4%)

Table 9. Comparison of the Viola-Jones algorithm results

Region	Images		
	Color	Grayscale	Our method
Faces	146(58.4%)	**185(74%)**	**246(98.4%)**
Nose	66(26.4%)	**69(27.6%)**	**206(83.8%)**
Eyes	0(0%)	1(0.004%)	**197(80.2%)**
Mouth	**70(28%)**	41(16.4%)	**157(63.7%)**

response. The next step is to segment the person's face using digital image processing techniques such as image binarization and morphological operators. After the interest region is determined, preprocessing is performed using a combination of linguistic modifiers and histogram adjustment to enhance the face in the image.

In this part, the Viola-Jones algorithm is implemented to find the interest regions. Once they are found, the algorithm crops the image and searches for an optimal threshold to binarize this region and segment it completely. Using this method, 250 effective images were collected, of which 246 complete faces were segmented, with an 98.4% effectiveness. Five different combinations of linguistic modifiers were used in the preprocessing, yielding different results. To evaluate algorithm performance, the average of the five tests performed in each region was obtained. For the nasal region, the average was 83.8%, which corresponds to an average of 206 recognized noses. For the mouth region, the average was 63.7% with a total of 157 mouths recognized. For the eye region, the average was 80.2%, corresponding to 197 segmented pairs of eyes. To verify whether the proposed method was effective, two comparative studies were performed. First, the proposed method was compared with the Viola-Jones algorithm performance by analyzing the thermal images in color and grayscale. Test results show that the proposed method outperforms the Viola-Jones algorithm by far and achieves better performance values for the face, nose, eyes, and mouth regions. The second study was conducted by comparing the classification rate with different methods described in the literature. In this study, we showed that the proposed method outperforms the authors with whom it was compared to. It should be noted that the work with which it has been compared to searched only for the face within the thermal image. In our case, we searched for all parts of the face that were of interest, such as the nose, mouth, and eyes, which makes the method more effective when applying the Viola-Jones algorithm.

Table 10. Comparison of results with methods from the literature

Method	Region				Ref
	Faces	Nose	Eyes	Mouth	
Viola-jones	91%	–	–	–	[9]
Viola-jones	**93%**	–	–	–	[13]
Viola-Jones	78.99%	–	–	–	[17]
PM/CM-VJ	90.5%	–	–	–	[10]
Cross-examples+VJ	67%	–	–	–	[21]
VJ-HOG-Otsu's	82.33%	–	–	–	[2]
VJ-LBP-Otsu's	**94.66%**	–	–	–	[2]
VJ-Haar-like-Otsu's	88.66%	–	–	–	[2]
Our method	**98.4%**	**83.8%**	**80.2%**	**63.7%**	

6 Goals and Future Work

In future work, the method used in this work will be modified. First, the environment in which the images are acquired will be better controlled. Other DIP techniques will be applied to improve the segmentation. Image enhancement will be done by trying more linguistic modifiers' combinations and other ways to adjust the image histogram. The goal is to maximize the accuracy of the Viola-Jones

algorithm and, on the other hand, increase the number of images to be used.

References

1. **Ashrafi, R., Azarbayjani, M., Tabkhi, H. (2022).** Charlotte-thermalface: A fully annotated thermal infrared face dataset with various environmental conditions and distances. Infrared Physics Technology, Vol. 124, pp. 104209. DOI: https://doi.org/10.1016/j.infrared.2022.104209.
2. **Basbrain, A. M., Gan, J. Q., Clark, A. (2017).** Accuracy enhancement of the viola-jones algorithm for thermal face detection. Intelligent Computing Methodologies: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part III 13, Springer, pp. 71–82.
3. **Blanco-Ruiz, M., Sainz-de Baranda, C., Gutiérrez-Martín, L., Romero-Perales, E., López-Ongil, C. (2020).** Emotion elicitation under audiovisual stimuli reception: Should artificial intelligence consider the gender perspective?. International Journal of Environmental Research and Public Health, Vol. 17, No. 22. DOI: 10.3390/ijerph17228534.
4. **Cao, Q., Qingge, L., Yang, P. (2021).** Performance analysis of otsu-based thresholding algorithms: a comparative study. Journal of Sensors, Vol. 2021, pp. 1–14.
5. **Fardian, F., Mawarpury, M., Munadi, K., Arnia, F. (2022).** Thermography for emotion recognition using deep learning in academic settings: A review. IEEE Access, Vol. 10, pp. 96476–96491. DOI: 10.1109/ACCESS.2022.3199736.
6. **Hassan, B. A., Dawood, F. A. A. (2023).** Facial image detection based on the viola-jones algorithm for gender recognition. International Journal of Nonlinear Analysis and Applications, Vol. 14, No. 1, pp. 1593–1599.
7. **He, S., Wang, S., Lan, W., Fu, H., Ji, Q. (2013).** Facial expression recognition using deep boltzmann machine from thermal infrared images. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 239–244. DOI: 10.1109/ACII.2013.46.
8. **IR-Fusion, F. T. (2023).** https://www.fluke.com/es-mx/search/fluke/?query=IR-Fusion%C2%AE. none.
9. **Kopaczka, M., Nestler, J., Merhof, D. (2017).** Face detection in thermal infrared images: A comparison of algorithm- and machine-learning-based approaches. Advanced Concepts for Intelligent Vision Systems, Springer International Publishing, Cham, pp. 518–529.
10. **Kwaśniewska, A., Rumiński, J. (2016).** Face detection in image sequences using a portable thermal camera. Proceedings of the 13th Quantitative Infrared Thermography Conference, pp. 4–8.
11. **Lin, C.-H., Wang, Z.-H., Jong, G.-J. (2020).** A de-identification face recognition using extracted thermal features based on deep learning. IEEE Sensors Journal, Vol. 20, No. 16, pp. 9510–9517. DOI: 10.1109/JSEN.2020.2986098.
12. **Liu, Z., Wang, S. (2011).** Emotion recognition using hidden markov models from facial temperature sequence. Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 240–247.
13. **Mekyska, J., Espinosa-Duró, V., Faundez-Zanuy, M. (2010).** Face segmentation: A comparison between visible and thermal images. 44th annual 2010 ieee international carnahan conference on security technology, IEEE, pp. 185–189.
14. **Michelini, Y., Acuña, I., Guzmán, J. I., Godoy, J. C. (2019).** LATEMO-E: A film database to elicit discrete emotions and evaluate emotional dimensions for Latin-Americans (supplementary material). Trends in Psychology, Vol. 27, No. 2, pp. 473–490. DOI: 10.6084/m9.figshare.5372782.v4.
15. **Mohamed, B., Haytam, H., Abdelhadi, F. (2021).** Applying fuzzy logic and neural

network in sentiment analysis for fake news detection: Case of covid-19. Studies in computational intelligence. DOI: 10.1007/978-3-030-90087-8$_{19}$.

16. **Pavez, V., Hermosilla, G., Pizarro, F., Fingerhuth, S., Yunge, D. (2022).** Thermal image generation for robust face recognition. Applied Sciences, Vol. 12, No. 1. DOI: 10.3390/app12010497.
17. **Reese, K., Zheng, Y., Elmaghraby, A. (2012).** A comparison of face detection algorithms in visible and thermal spectrums. Int'l Conf. on Advances in Computer Science and Application, pp. 49–53.
18. **Sai Prasanna, G., Pavani, K., Kumar Singh, M. (2022).** Spliced images detection by using viola-jones algorithms method. Materials Today: Proceedings, Vol. 51, pp. 924–927. DOI: https://doi.org/10.1016/j.matpr.2021.06.300. CMAE'21.
19. **Saravanan, C. (2010).** Color image to grayscale image conversion. 2010 Second International Conference on Computer Engineering and Applications, Vol. 2, pp. 196–199. DOI: 10.1109/ICCEA.2010.192.
20. **Tahseen, A.-J. A., Sotnik, S., Sinelnikova, T., Lyashenko, V. (2023).** Binarization methods in multimedia systems when recognizing license plates of cars. International Journal of Academic Engineering Research (IJAER), Vol. 7, pp. 1–9.
21. **Tran, H., Dong, C., Naghedolfeizi, M., Zeng, X. (2021).** Using cross-examples in viola-jones algorithm for thermal face detection. Proceedings of the 2021 ACM Southeast Conference, pp. 219–223.
22. **Yan, L., Cengiz, K., Sharma, A. (2021).** An improved image processing algorithm for automatic defect inspection in tft-lcd tcon. Nonlinear Engineering, Vol. 10, No. 1, pp. 293–303. DOI: doi:10.1515/nleng-2021-0023.
23. **Zhang, C., Shen, X., Cheng, H., Qian, Q. (2019).** Brain tumor segmentation based on hybrid clustering and morphological operations. International journal of biomedical imaging, Vol. 2019.

Article received on 10/06/2023; accepted on 13/01/2025.
**Corresponding author is Luis Alberto Hernández Montiel.*