Vision-based Systems Final Project

Alexandre Rocha
up201705277@fe.up.pt

Daniel Barros
up201704271@fe.up.pt

João Silva
up201704946@fe.up.pt

Ricardo Falcão
up201704220@fe.up.pt

I. Methods

A. Task1

1) Approach Pipeline: The goal of this task was to detect faces in a given image. The algorithm receives the raw image as input and outputs it with the detected faces' bounding boxes, signaling the *ROIs*. The result can later be used for mask detection, in *Task 2*.

2) Pre-Processing: The proposed algorithm begins by submitting the image to a series of filtering operations in order to improve the chance of finding skin in a posterior phase of the program and, therefore, increase the probability of generating a correct face detection.

To achieve such objective, a small Gaussian Filter is applied to the image. The result is a smoothed picture that will have a positive influence at this early stage.

Next, the image undergoes a lighting compensation operation. This culminates in a now color balanced picture, with less bright variations and, subsequently, better suited for the operations to follow.

In order to pass a cleaner input to the skin detection mechanism, a *K-Means Segmentation* is applied. Through an empirical observation, a series of clusters were analysed and found to have a strong relation with unwanted information. Their values provided inferior and superior limits that were used in constructing a mask with only rejected objects. This is then used in removing the excess parts from the previous step's output. They consist of background portions, hair and, mainly, clothes.

Afterwards, the image is ready to suffer a selective threshold segmentation that removes almost all non-skin areas. A lot of different color spaces were used but, in the end, only the *RGB* one proved to be effective in a consistent manner. The set values were initially based on [2] and adapted to better integrate the algorithm. They were modified, in general, to allow a greater tolerance, made possible by the previous use of clustering.

Once this stage is over, a set of operations is applied to end the preparation of the image, before entering the next phase. It consists of a morphological closing, a filling of the image's objects' holes and a purge of small regions according to the objects relative size.

3) Iterative Face Extraction: At this point, the now very filtered image enters a loop where, for a small predefined number of iterations, a developed function checks region by region if a face is currently in the input.

To accomplish this goal, the regions are analysed and a considerate number of features is extracted from each one. These properties are then submitted trough a vetting process. Only if all the chosen conditions for being considered a face are met, is the labeled region signaled as a face. If such finding happens, the associated bounding box is stored in an array and the region is removed from the image in that iteration. This way a detected face cannot suffer any further changes.

The regions that fail the selection sequence remain in the image to be subjected to further processing. An edge detection is applied, along with morphology operations and light segmentations of regions considered small when compared to their neighbours.

If, at the end of the loop, the remaining objects, after the processing, are not approved by the face extraction operation by becoming an acceptable face, then they are discarded.

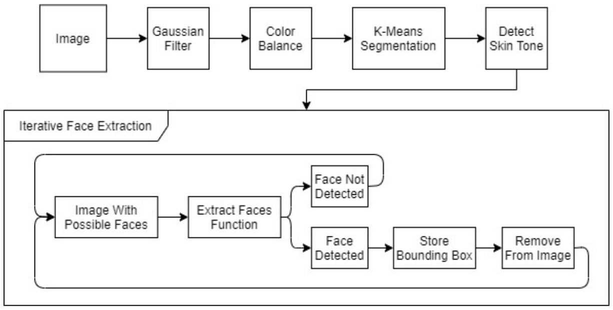

Figure 1. Face Detection Algorithm Diagram

B. Task2

1) Approach Pipeline: The goal of this task was to detect if a certain person is wearing a face mask and, if so, there is the need to understand if it is correctly worn - one can define correctly worn as a mask placed in such a way that is covering both nose and mouth. It is also important to note that, as an input, the algorithm receives one image and the existing faces *ROIs*, i.e., the output of the previous *Task 1*.

Given those inputs, the proposed algorithm starts by constructing a binary mask responsible for cancelling surrounding non-skin areas, based on the selective threshold segmentation developed in *Task 1*. Afterwards, a convex hull of the face pixels was considered. However, in some cases, it proved to be worse, sometimes even deleting desired facial features. Next, it detects if one's lips are visible. If they are, it means that no face mask is worn. However, if they're not, there is still the need to detect the nostrils, in order to understand the placement of the mask. An overview of the algorithm is presented in Figure 2.

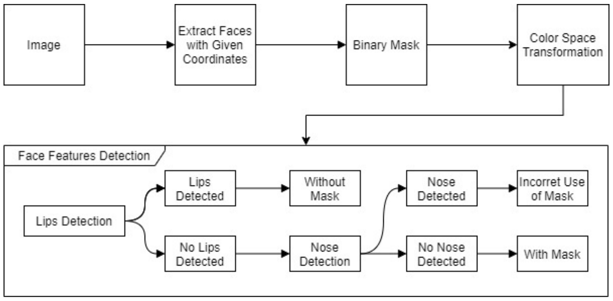

Figure 2. Mask Detection Algorithm Diagram

2) Lips Detection: Regarding this subject, the authors in [1] present a method to highlight lips through the enhancement of the red color's tone. Using the color space YC_bC_r, the red tone is denoted by higher intensities of the C_r component, as well as lower intensity of the C_b one, when comparing with skin tone.

Given this work's accuracy on detecting one's lips, the algorithm uses the same mouth map:

$$MouthMap = C_r^2(C_r^2 - \eta C_r/C_b)^2$$
$$\eta = 0.95\frac{\frac{1}{n}\sum_{(x,y)\in BM} C_r(x,y)^2}{\frac{1}{n}\sum_{(x,y)\in BM} C_r(x,y)/C_b(x,y)}$$

where BM represents the binary mask, previously constructed.

To improve the lips' contrast in relation to other face zones, a morphological top-hat transform and a dilation are used on top of the result of $MouthMap$. In addition, to increase accuracy, a red color mask constructed from the input image in the HSV color space is applied to guarantee that only pixels with a strong red component are considered. The image is then binarized through an empirically determined threshold, with the assumption that, if lips are visible, these must now be the brightest object in the image.

Finally, it is necessary to perform an image analysis regarding the number of regions found and their properties. At this point, to prevent noise detection, only the bottom half of the image is considered, once the lips must, by all means, be in this zone. The following conditions are sequentially checked in order to conclude if lips are visible. If one fails, the others are ignored and the algorithm considers that it couldn't detect any lips.

- Number of regions: Only one region must be detected.
- Orientation: Considering that one's face is, in the worst case, 45 degrees leaned, any leaning greater than this value is discarded.
- Circularity: Any object that has not the minimum circularity characteristic (e.g. a long line, a concave object, etc.) is not considered.
- Centroid: Detected objects right next to the image margins are also discarded, since it most likely will be noise.

3) Nostrils Detection: Once again, the algorithm is based in the work presented in [1]. Here, the authors present not only a sequence of operations to enhance one's mouth, but also one to enhance the eyes, with high accuracy. This algorithm uses the basis of detecting eyes and applies a similar method to the detection of nostrils, as these two face features share common properties as a dark center and a circular brighter surrounding, with a sharp variation. As presented in the referred work, it's possible to develop gray scale morphological operators to emphasize those brighter and darker pixels in the Y component, when in YC_bC_r color space. Hence, the creation of the Y eye map consists in the following equation:

$$EyeMapY = \frac{Y(x,y) \oplus g_\sigma(x,y)}{Y(x,y) \ominus g_\sigma(x,y) + 1}$$

where the gray scale operators $\oplus$ and $\ominus$ are dilation and erosion respectively. To enhance the nose region in contrast to other surrounding face zones, a morphological top-hat transform and opening are applied. At this point, if nostrils are visible, these must be the brightest objects in the image, together with the eyes. Thus, to isolate them, an empirically determined threshold is applied.

It is important to refer that, although the authors have considered the construction of a chromatic C_bC_r eye map to lately multiply by the Y one, this algorithm discards the former for the reason that this operation does not enhance the nostrils as desired, sometimes even deteriorating it.

To finish, a projection algorithm is implemented in order to get the pixel coordinates of the vertical projection maximums. It is expected that these coordinates span the respective pixel within the region between the eyes and the mouth. Therefore, if the nostrils are not covered by the face mask, at least one maximum of the vertical projection will belong to this region.

C. Task3

1) Approach Pipeline: The goal of this task was to combine the results obtained from *Task1* and *Task2*. The algorithm receives the raw image as an input and outputs it with the detected faces' bounding boxes, as well as the evaluation on whether or not the detected person has a correctly placed mask. Having in mind the intention to implement this task, the code was developed in a modular pattern, so it could be reused as it was needed. After this constraint, the implementation of the *Task3* algorithm proved to be extremely simple.

II. Results and Discussion

A. Task1

Regarding the final face detection path taken, one can conclude that the algorithm is very effective in recognizing human faces in relatively easy situations. However, the more crowded the image gets as well as the smaller the *ROIs* become, it's noticeable that the mentioned effectiveness starts to decline. This is due to certain decisions made along the process of designing such algorithm.

In a very early stage it was decided that the best global approach would be focused on skin detection so, in order to build a consistent mechanism around that, a large variety of taught methods was tested. These approaches will be discussed next, phase by phase.

1) Pre-Processing: In order to pass a consistent input to the face extraction phase, different pre-processing techniques were implement along the way.

The first one consisted of a Superpixels algorithm that smoothed the image according to a region local mean, for each region. This was found to be very effective in images were the *ROIs* were separated and had large surfaces, but it was rapidly considered ineffective for smaller and closer faces, as well as images with non distinct backgrounds.

In replacement, a less aggressive Gaussian Filter, with a small *sigma*, removed some of the image's sharpness and a color balance sequence normalized the intensities of each *RGB* channel, but sometimes added blue tones to the images. That required re-calibration of the thresholds. These methods were able to generate a solid input for the skin detection mechanism. However, these limits were too selective and proved to be inefficient for some skin tones.

In order to have a more tolerant discrimination, a *K-Means segmentation* was implemented just after the color balance. This was initially intended to help the skin detection directly but it became clear that only with a very large number of clusters could that be useful. This was caused by the difficulty of picking the right clusters associated with skin tones, consequence of the large range of intensities presented that created unpredictability. After a closer examination, a pattern emerged. There were always some clusters, withing the same range of center values, that were associated with undesired parts of the image. With that information, the unwanted excess was removed and that created room for a larger tolerance in the *RGB* thresholds. The skin detection mechanism became very effective, but also slower due to the heavy processing required in the clustering algorithm.

For most of the project duration, the YC_bC_r color space, mentioned in many papers, was used along with the *RGB* one. This, however, became obsolete towards the end and was removed for inflicting some damage during the images' filtering.

Finally, some final minor operations had to be added. To fix the skin fragmentation, a morphological closing with a medium sized disk, followed by a filling of the holes and a removal of small fragments were applied.

It's important to mention that a distance transform method was introduced to separate faces that were very close to each other. This presented, in some cases, clear centers, with high intensities, that along side with region growing could be used to separate the *ROIs*. This method was very interesting approach for the medium/difficult images in the data-set. But, ultimately, was still very inconsistent and, therefore, was removed from the final algorithm.

2) Iterative Face Extraction: The previous stage's output presented images mostly with human faces and limbs. Sometimes, unrecognizable in the first iteration for being connected. In order to select only clear faces, a labeling of the regions is implemented using an 8-connected association. Each of these regions, for a small number of predefined iterations, is then submitted to a strict vetting process made by a set of restrictions. These conditions were created based on the analyses of the region's properties found in True Positives, acquired along the experiments made. An excel document was created to retrieve inferior and superior limits to best calibrate each feature interpretation, based on the statistics created. Some of the more important features are eccentricity, width to height ratio and relative orientation. It was also implemented a sequence that eliminates regions based on their relative size when compared to their neighbours. This saves the maximum size present and makes a decision according to it.

If an object passes the tests, it's bounding box is stored and added to the original image's red channel, therefore signaling a detection. This region is then removed from the iterations to prevent further unwanted processing.

The objects remaining will suffer an *Canny* edge detection, followed by a small morphological closing, a filling, a small erosion and a purging of small areas in relation to the image. These operations allow the algorithm to separate regions, increasing the probability of detecting a face.

After testing the provided image database, these were the results:

Table I
EASY SAMPLES
($R = 90.9\%$, $P = 76.9\%$, $F_1 = 83.3\%$)

		Predicted Label	
		Class 1	Class 2
True Label	Class 1	20	2
	Class 2	6	-

Table II
MEDIUM SAMPLES
($R = 23.9\%$, $P = 57.9\%$, $F_1 = 33.8\%$)

		Predicted Label	
		Class 1	Class 2
True Label	Class 1	11	35
	Class 2	8	-

Table III
HARD SAMPLES
($R = 1.8\%$, $P = 50\%$, $F_1 = 3.4\%$)

		Predicted Label	
		Class 1	Class 2
True Label	Class 1	3	167
	Class 2	3	-

While the medium and the hard difficulty samples proved to be quite difficult to achieve good results, the easy samples turned out really good. Although not perfect, and some interesting conclusions were drawn on this task, and we believe that it is very acceptable for a first approach of this problem.

B. Task2

The performance of this algorithm must be evaluated according to the type of input it receives. Naturally, any image can be passed as an argument. Although, there are huge differences between images where faces are sideways and far from the camera focal point, and images where faces have an acceptable dimension, with higher resolution and ideally have both eyes visible. In the former, as expected, the algorithm has a lower performance due to lack of resolution. Here, face features does not follow any consistent pattern of shape, texture, or even color. In contrast, in the latter, nostrils are round and well defined, the red color in lips regions is well detected and face masks, if any, represent a significant area with respect to total face area.

This work is mainly focused on detecting face masks where faces have acceptable resolution. Therefore, it is appropriate to start by evaluating the performance of this cases, proving the higher performance previously described. The confusion matrix of the first 11 images from the ground truth set provided, which follow these characteristics, can be analyzed in Table II-B.

Table IV

		Predicted Label		
		Class 1	Class 2	Class 3
True Label	Class 1	4	0	0
	Class 2	0	2	1
	Class 3	0	1	3

Considering the following correspondences:

- Class 1 - Without mask
- Class 2 - With mask well placed
- Class 3 - With mask worn incorrectly

The calculated metrics are shown in the following table:

Table V

	Class 1	Class 2	Class 3
Recall	100%	66.67%	75%
Precision	100%	66.67%	75%
F-Measure	100%	66.67%	75%

However, it is also of utmost importance to evaluate an overall result, with different images, with different face poses, skin colors, face mask colors and textures, etc. The complete confusion matrix of the total ground-truth set provided is now presented in Table .

The calculated metrics are shown in the following table:

Table VI

		Predicted Label		
		Class 1	Class 2	Class 3
True Label	Class 1	5	29	5
	Class 2	1	58	14
	Class 3	2	11	7

Table VII

	Class 1	Class 2	Class 3
Recall	12.82%	79.45%	35.00%
Precision	62.50%	59.18%	26.92%
F-Measure	21.28%	67.84%	30.43%

Clearly, the algorithm performs quite better in the first cases. However, given the data set difficulty, with low resolution images, the overall result is satisfactory. With some improvements, mainly in nostrils detection, it is highly probable that this algorithm would improve significantly.

C. Task3

After implementing the algorithm for the task3, and testing the provided image database, some issues were discovered because of the decisions made on the development of the previous tasks. This was mainly because of the characteristics that were defined in the face evaluation of the task 1, that made it really hard to detect not only smaller faces, but also faces that were half-hidden (which is the case of a person wearing a face mask).

If the task1 could detect the faces correctly, success rates similar to the ones obtained in the task2 would be expected.

Unfortunately, not a single face was detected from the provided image database, so no results can be presented.

REFERENCES

[1] Hsu, Rein-Lien & Abdel-Mottaleb, Mohamed & Jain, Anil. (2002). Face detection in color images. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1. 696 - 706. 10.1109/34.1000242.

[2] Kovac, J. & Peer, Peter & Solina, Franc. (2003). Human skin color clustering for face detection. EUROCON. 2. 144 - 148 vol.2. 10.1109/EURCON.2003.1248169.