Face recognition using convolutional neural networks and metadata in a feature fusion model

ABSTRACT

Recent advances in science and technology are raising ever-increasing security issues. In response, traditional authentication systems based on knowledge or possession have been developed, but these soon came up against limitations in terms of security and practicality. To overcome these limitations, other systems based on the individual's unique characteristics, known as biometric modalities, were developed. Of the various ways of improving the performance of biometric systems, feature fusion and the joint use of a pure biometric modality and a soft biometric modality (multi-origin biometrics) are highly promising. Unfortunately, however, we note a virtual absence of multi-origin systems in a feature fusion strategy. For our work, we therefore set out to design such a multi-origin system fusing facial features and skin color. Using OpenCV (Open Computer Vision) and Python, we extracted facial features and merged them with skin color to characterize each individual. The HOG (Histogram of Oriented Gradients) algorithm was used for face detection, and Google's deep neural network for encoding. For skin color, segmentation in the HSV (Hue, Saturation, Value) color space enabled us to isolate the skin in each image, and thanks to the k-means algorithm we had detected the dominant skin colors. The system designed in this way enabled us to go from 81.8% as a TR (Recognition Rate) with the face alone to 86.8% after fusion for a TFA (False Acceptance Rate) set at 0.1% and from 0.6% as a TEE (Equal Error Rate) to 0.55%.

Keywords: multibiometrics, face, skin color, feature fusion.

1. INTRODUCTION

Today's world, more than in the past, is dominated by scientific and technological progress, judging by the many achievements to date. At the heart of all these advances has always been the eternal question of security. Who has the right to access which resources? How can we be sure of a claimant's identity? How can each individual be uniquely identified? So many questions showing just how crucial security is. And with security threats on the increase, protecting individuals and institutions has become a top priority.

Existing traditional systems (knowledge-based and token-based) soon proved to be limited in terms of security. For improved security, biometrics appeared to be a promising solution, using unique and irreversible characteristics of the individual for identification purposes. However, single-mode biometrics (the use of a single characteristic modality for identification purposes) revealed its own limitations. Multibiometrics was developed to overcome the shortcomings of single-mode biometric systems, such as the public unacceptability of certain modalities, the absence of certain modalities in certain individuals, performance limitations and vulnerability to identity theft [1]. Multi-biometrics can reduce the impact of these problems, but cannot entirely eradicate them, and new challenges have emerged in the use of multi-biometric systems, such as ease and cost of implementation.

In terms of performance, biometric information fusion is a key aspect for improving recognition rates in multi-biometric systems. There are four (04) levels of fusion in biometric systems. Score-level fusion has been the most widely studied in the literature, due to its ease of implementation. But several authors point out that scores provide only a limited amount of information [2]-[4]. If we wish to further improve fusion performance, we would need to move into a more information-rich space, such as that of features [5].

Unfortunately, we note the non-existence of multi-origin systems using pure and soft biometric modalities in a feature fusion strategy.

In this paper, we propose a method for fusing facial features and skin color, taking into account the fact that the face is a modality rich in information about an individual's morphological traits, and that skin color is a discriminating soft biometric that can be acquired simultaneously with the face and without contact. Once our system has been designed, we evaluate its performance and compare the results obtained with those of existing methods.

2. MATERIAL AND METHODS

2.1. Materials

2.1.1. Hardware

In carrying out this work, we obviously made use of a computer to implement our solution. The characteristics of this computer are:

- Model: HP EliteBook 840 G1
- Processor: Intel® Core™ i5-4300U
- Speed: 2.5 GHz
- Memory: 256GB SSD
- Ram: 8GB
- Intel® HD Graphics Family
- Operating system: Windows 10 Professional 64-bit

2.1.2. Software

The software package includes programming languages, software and useful libraries. It includes:

- Python 3.8;
- VS Code with integrated Jupyter Notebook;
- Open CV;
- Numpy, matplotlib, Scikit-learn.

2.1.3. Database

We used Casia Face V5, a facial image database developed by the Institute of Automation, Chinese Academy of Sciences (CASIA). It includes 2.500 images of 500 different people, 05 images per person. Figure 1 below shows an overview of this database. All images are 16-bit color BMP files, with an image resolution of 640*480. Typical intra-class variations (lighting, pose, expression, glasses, imaging distance, capture angle, etc.) make this a robust database for facial recognition research.

Casia Face V5 is widely used for training and evaluating facial recognition models, as it covers a wide variety of illumination, expression, gender and age conditions. Casia Face V5

volunteers include graduate students, workers, etc. In our work, we use 03 images per person for enrolment and 02 images for testing.

2.1.4. Color spaces

Color spaces are color coding systems used to describe and manipulate colors in digital images. We present here the two color spaces used in our work: RGB and HSV.

- The **RGB (Red, Green, Blue)** color model is one of the most commonly used color spaces in color representation. It is based on the principle that any visible color can be obtained by mixing different quantities of red, green and blue light. The RGB color space is a three-dimensional system in which each axis represents the intensity of the corresponding component (red, green or blue). Component values generally range from 0 to 255, where 0 represents the absence of that component and 255 represents the maximum intensity. By normalizing these values, they can be represented in a range from 0 to 1, where 0 represents the absence of the component and 1 represents the maximum intensity. The vertices of each axis, i.e. (1,0,0) represent pure red, (0,1,0) represents pure green and (0,0,1) represents pure blue. The input image in this color space is defined as follows:

  $X_{ij} = [r_{ij}, g_{ij}, b_{ij}]$ (1)

  Here, (ij) is the pixel coordinate, r_{ij}, is the value of the R component, g_{ij}, is the value of the V component, and b_{ij}, is the value of the B component. In RGB color space, the primary colors are red, green and blue, represented respectively by the vectors (1,0,0),
  (0,1,0) and (0,0,1). Mixing these three primary colors produces all the other colors in the RGB space. For example, equal mixing of red and green produces yellow, while equal mixing of red and blue produces magenta. By adjusting the intensities of the three components, it's possible to create an infinite range of colors. It's important to note that the RGB color space is not perfectly suited to all situations. For example, it does not perfectly represent certain subtle colors, and may present limitations in the accurate representation of real colors. In such cases, other color spaces, such as CIE Lab (Luminance, a, b), YCrCb (Luminance, Chrominance Red, Chrominance Blue) or HSV (Hue, Saturation, Value) as required, are often used for more accurate and consistent color representation.
- **HSV (Hue, Saturation, Value)** color space: This is a cylindrical system that defines colors using hue (hue), saturation (saturation) and luminance (value). This space is particularly useful for color-based operations, such as the selection of specific colors. It has been widely used in biometrics for face identification. For example, in the recent article by Nguyen et al. [6] the authors proposed a face detection system based on HSV space. The proposed method achieved a detection rate of 96.25%. It was also exploited by Sobabe et al. [7] for the extraction of skin color features to authenticate individuals.

2.2. Methods

Facial recognition in our system:

- Face detection: the HOG descriptor

  The HOG (Histogram of Oriented Gradients) descriptor is an object detection method based on the analysis of image gradients. A gradient is a measure of the variation in pixel intensity in a given direction. The HOG descriptor calculates a histogram of gradient orientations in a region of the image, called a cell. By

combining the histograms of several cells, we obtain a compact, discriminating representation of an object's shape and appearance.

The HOG descriptor was proposed by Dalal and Triggs [8] in 2005 for the detection of pedestrians in images. Since then, it has been successfully applied to other object detection tasks, such as face, car and bicycle recognition. The HOG descriptor has several advantages over other object detection methods:

- It is robust to variations in lighting, color and texture, as it is based on gradients, which are more stable than pixel intensities.
- It captures the contours and shapes of objects, which are discriminating features for object detection.
- It is invariant to local geometric transformations, such as rotations or translations, because it uses normalized histograms that do not depend on the absolute position of pixels.
- It is efficient to calculate, using simple operations such as filtering, histogram calculation and normalization.

This method is commonly used in object recognition for pattern detection. It captures the local features of an image by analyzing the orientation gradients in the image.
The HOG descriptor thus captures important image information, such as contours, shapes and textures, while ignoring unnecessary information, such as color or overall intensity. The HOG descriptor therefore provides a vector that characterizes the image according to the spatial distribution of gradient orientations.

– **Face encoding: FaceNet's convolutional neural network**

FaceNet is a deep learning-based face recognition system developed by Google researchers in 2015 that offers an innovative solution to the dimensionality problem of vector representations of faces while maintaining high performance. It uses a convolutional neural network (CNN - Convolutive Neuronal Network) to learn a vector representation of faces, called embedding or integration, which captures the distinctive features of each individual. The similarity between two integrations can be measured by Euclidean distance or cosine distance, enabling faces to be compared and identified.

This neural network does not need to be trained for each individual to be recognized, which means it can handle an unlimited number of people without having to re-train the model. It has outperformed other methods on several public databases, such as Labeled Faces in the Wild (LFW) or YouTube Faces (YTF).

For face detection in our system, we use the HOG descriptor as described. Once the face has been detected, we crop the initial image. We then apply an alignment to this new image to prevent faces from being turned in different directions and to impose a fixed direction on all faces processed by the system. The resulting image is the one we'll encode next. For encoding, we use FaceNet's convolutional neural network. This network provides us with a vector of 128 values, which we use to classify the individuals.

– **Segmentation of skin color by thresholding, feature extraction and adapted fusion**

For an individual, skin color varies from one region of the face to another and from one image to another, depending on lighting conditions. Thus, for the same person,

the R, G and B components of the dominant skin color may vary. It is therefore not possible to record a fixed color value for an individual. We therefore determine a confidence interval for each of the R, G and B components. The procedure used here is taken from the work of Sobabe et al. [7].

For our database, we use the 03 images available per individual and proceed as follows: For each image, the previously extracted face is subjected to a segmentation function in HSV color space in order to retain only the skin pixels. In line with the work of Sobabe et al. [7] for the H, S and V components, the values of the lower and upper bounds to be considered as skin color are respectively: [0, 48, 80] and [20, 255, 255] respectively. Empirical thresholding enables us to keep the pixels representing skin color and eliminate the others. A new image is thus obtained.

We apply the k-means algorithm to this new image to extract the dominant colors. We vary the number of clusters from 3 to 40 to determine the number that would give the best results. In our case, the best results are obtained with 20 clusters. Next, we select the two dominant colors and for each of these two colors, the R, G and B components of the color space are isolated. These values are stored in three different tables for the 03 images: one for the R component, one for the G component and one for the B component. Each of the three tables thus contains 06 values, based on 2 dominant colors multiplied by the 03 images.

Next, the k-means algorithm is applied again (with a cluster number equal to 2) to each of the three tables to produce two classes of values per table. The first class represents low values and the second-high values. The centroid of each class is then selected. This provides each class with a representative value characterizing the class. In this way, each table gives two values corresponding to useful information on each component of the RGB color space. These two values were used to define the confidence interval required for each of the R, G and B components.

These 03 confidence intervals are stored in the database for this individual. We therefore have a total of 6 values recorded in the model for each individual.

The average value of each of these three classes allows us to obtain 03 values constituting the features derived from skin color, which are merged by concatenation with the face feature vector to form the vector characterizing this image.

For each test image, once the face has been extracted and segmented, we apply the k-means algorithm, retaining only the dominant color. During the comparison, we check whether at least 2 out of 3 of the components of this color belong to the confidence interval retained in the reference database during enrolment. If so, we add these 03 components to the 128 face values, and compare the new vector with the one obtained during enrolment. If no, we judge that skin color characteristics will not be able to improve the results obtained with the face alone; we therefore carry out the comparison with the 128 face values.

An explicit description of the metadata analysis algorithm is as follows.

As the metadata (skin color) taken on its own does not allow us to establish the identity of an individual, as described above, we have proposed an algorithm for taking it into account in our system. In this algorithm, we first check whether the

content of each RGB component of the dominant color of the input image belongs to the range constituted by the contents of these same components of the enrolment or reference images.

Let A_1, A_2 and A_3 respectively, be the intervals constituted by the R, G and B contents of the two dominant colors of all the enrolment images of individual i. Let us denote by x, $y$and z respectively, the R, G and B contents of the dominant color of the input image of the individual i

$$If(x \in A_1\ AND\ y \in A_2)OR(x \in A_1\ AND\ z \in A_3)\ OR(y \in A_2\ AND\ z \in A_3)(2)$$

then the metadata is taken into account and we concatenate the characteristics of the metadata with those of the pure biometric modality before moving on to the decision module. Otherwise, the metadata is ignored to avoid degrading the results initially obtained with the face.

- Comparison module

For our comparison module, we explored several methods: Euclidean distance, Manhattan distance, Chebyshev distance, Spearman distance, cosine similarity and correlation. So, for each person, we compare the characteristics of each of the 02 test images with those of each reference image, using these comparison methods. We proceed to a majority vote: the face on a test image is recognized as a client if it matches at least 02 of the 03 reference images. To determine the best threshold for our system, we vary the threshold from 0 to 1 in steps of 0.001. We set the False Rejection Rate at 0.1% to enable a better comparison between our similarity measurement methods on the one hand, and between our results and previous work on the other.

3. RESULTS AND DISCUSSION

3.1 Results obtained with the face only

We had set up a reference image base of 03 images per person and a test image base of 02 images per person. Each of these images is subjected to the HOG algorithm described above to detect and extract the individual's face. This face is passed as input to a pre-trained deep convolutional neural network, which generates 128 values as features. We then compare the integration of each test image with the integrations of the 03 test images per person. We consider a client if the test image matches at least 2/3 of the reference images per person. In this way, we obtain the results shown in Table 1.

Table 1. Results obtained with the face only

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	81.8%	0.6%	98.9%
Euclidean	77.7%	0.6%	98.875%
Correlation	1.125%	8.0%	94.55%
Cosinus	70.4%	2.15%	98.753%
Chebyshev	75.675%	6.0%	98.133%
Spearman	80.6%	4.2%	98.523%

From this table, we can see that all the comparison methods except correlation offer us very high Recognition Rates (over 70%) for a False Reject Rate set at 0.1%, very high AUCs (over 98%) and relatively low Equal Error Rates. But the thresholds for a FAR = 0.1% differ widely. We recall that the lower the EER, the better the system's performance. Based on this criterion, we have selected the Euclidean and Manhattan distances from these 06 comparison methods, since they give the lowest EER (i.e. EER = 0.6%). Furthermore, while the Manhattan distance achieves an AUC of 98.9%, the Euclidean distance only achieves 98.875%. We can therefore deduce that the Manhattan distance outperforms the Euclidean distance as a comparison method for our data. This can also be clearly seen in the Recognition Rate achieved by each of these two distances for a FAR = 0.1%: the Manhattan distance wins out at 81.8% against 77.7%.

The performance curves for the Manhattan distance are shown below.

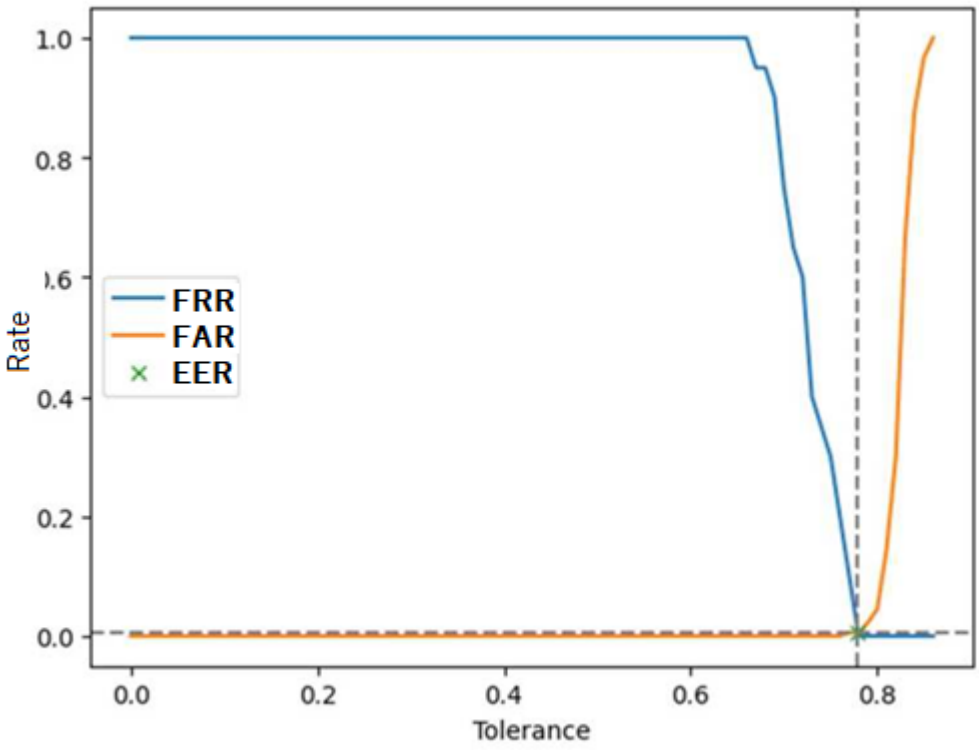

Fig. 1. False Reject Rate (FRR) and False Acceptance Rate (FAR) curves as a function of system tolerance

This figure shows the False Rejection Rate (FRR) and False Acceptance Rate (FAR) curves as a function of system tolerance for the use of the face modality only with Manhattan distance. It shows that initially (for zero tolerance), FRR = 1 and FAR = 0, and that as tolerance increases, FRR decreases while FAR remains almost zero. The two curves cross at the point (0.2382; 0.006). The FAR then increases until it reaches 1, while the FRR becomes zero. We can clearly see this evolution in the figure above.

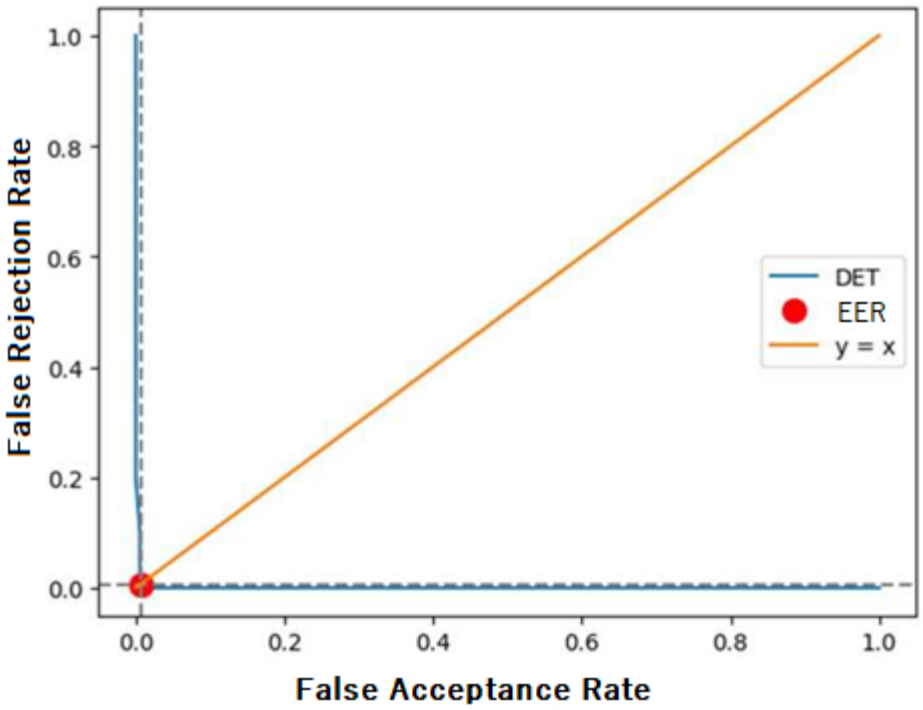

Fig. 2. DET curve of the system using the face modality only with the Manhattan similarity method.

This curve represents FRR as a function of FAR, and shows quite clearly the evolution of customers rejected by the system compared to imposters accepted.

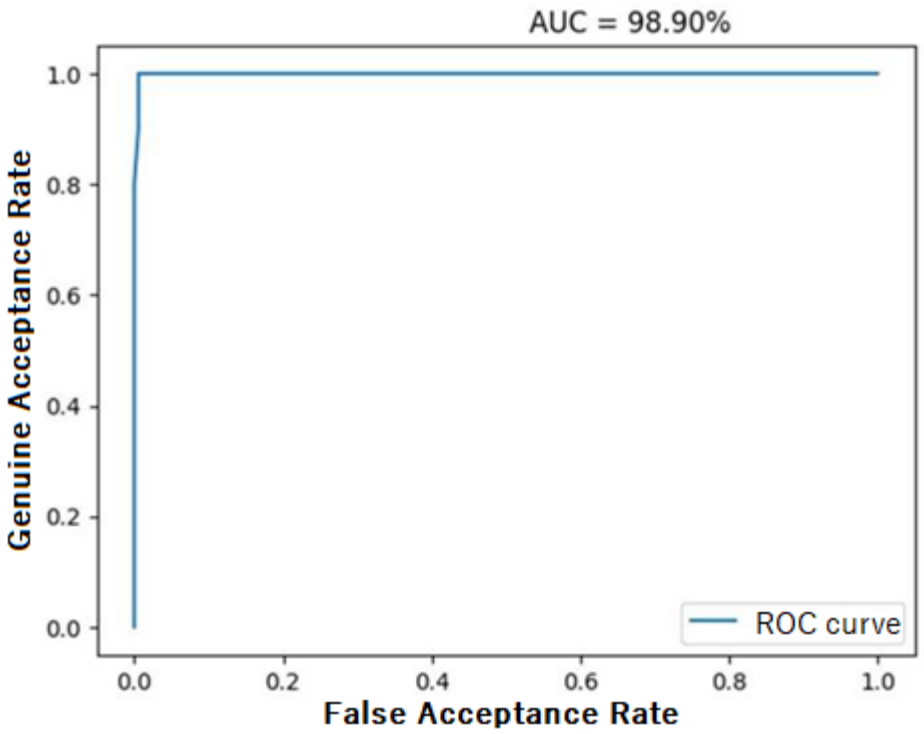

Fig. 3. ROC curve for the system using the face-only modality with the Manhattan similarity method.

The figure above shows the Receiver Operating Characteristic (ROC) curve for the facial recognition system. This curve is commonly used to evaluate the performance of a biometricsystem, as it provides a compact representation of performance for different parameter configurations in a single curve, enabling objective comparison between different systems. The curve represents the relationship between legitimate and false acceptance rates for different decision threshold values. The AUC (Area Under Curve) of the ROC curve is a measure of system efficiency, representing the area under the curve. An AUC value of 1 indicates perfect performance, while a value of 0 indicates zero performance. For our facial recognition system, the AUC is 0.989, showing that the system is not yet perfect, but is capable of recognizing and authenticating individuals.

3.2. Presentation of the results of facial recognition fused with skin color

Integrating skin color into the recognition process gives us the following results:

- For a number of clusters = 3

Table 2. Results of facial recognition merged with skin color for 3 clusters

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	62.7%	15.0%	89.457%
Euclidean	58.6%	25.0%	79.04%
Correlation	0.18%	37.4%	67.776%
Cosinus	0.132%	42.65%	63.318%
Chebyshev	40.4%	30.0%	72.964%
Spearman	72.7%	4.65%	98.261%

- For a number of clusters = 5

Table 3. Facial recognition results merged with skin color for 5 clusters

Distance type	TR at TFA = 0.001	TEE	AUC
Manhattan	61.8%	25.0%	81.589%
Euclidean	66.8%	25.0%	78.839%
Correlation	0.159%	39.0%	65.971%
Cosinus	0.135%	35.6%	67.826%
Chebyshev	40.51%	30.0%	72.658%
Spearman	70.6%	8.3%	97.244%

- For a number of clusters = 10

Table 4. Results of face recognition merged with skin color for 10 clusters

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	81.8%	5.0%	94.05%
Euclidean	77.7%	5.0%	93.988%
Correlation	0.095%	48.4%	60.108%
Cosinus	0.095%	47.2%	61.905%
Chebyshev	70.321%	7.0%	93.114%
Spearman	51.8%	9.7%	94.837%

- For a number of clusters = 20

Table 5. Results of facial recognition merged with skin color for 20 clusters

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	86.8%	0.55%	98.925%
Euclidean	78.6%	1.0%	98.865%
Correlation	0.12%	45.4%	60.582%
Cosinus	0.12%	45.4%	61.032%
Chebyshev	70.386%	6.3%	97.83%
Spearman	18.6%	14.0%	93.918%

- For a number of clusters = 30

Table 6. Results of facial recognition merged with skin color for 30 clusters

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	86.8%	0.55%	98.925%
Euclidean	78.6%	1.0%	98.865%
Correlation	0.118%	46.05%	61.378%
Cosinus	0.118%	46.05%	62.05%
Chebyshev	70.386%	6.3%	97.83%
Spearman	20.368%	25.1%	84.095%

- For a number of clusters = 40

Table 7. Results of facial recognition merged with skin color for 40 clusters

Distance function	GAR at FAR = 0.1%	EER	AUC
Manhattan	86.8%	0.55%	98.925%
Euclidean	78.6%	1.0%	98.865%
Correlation	0.131%	43.2%	62.105%
Cosinus	0.131%	43.2%	62.255%
Chebyshev	70.386%	6.3%	97.83%
Spearman	4.5%	27.6%	81.043%

From these tables, we can see that the adaptive fusion of facial features and skin color leads to a degradation in system performance for all comparison methods with cluster numbers below 20. For a number of clusters greater than or equal to 20, we note that the EER drops only for the Manhattan distance, compared with the EER obtained with this same distance before feature fusion. Results deteriorate for all other comparison methods, but improve for the Manhattan distance. The improvement here is in terms of Recognition Rate and threshold at FAR = 0.1% as you can see in Table 5 where we went from 0.6% to 0.55% as EER, from 81.8% as GAR for face only to 86.8% and from 98.9% to 98.925% as AUC. These performances are illustrated in the figures below:

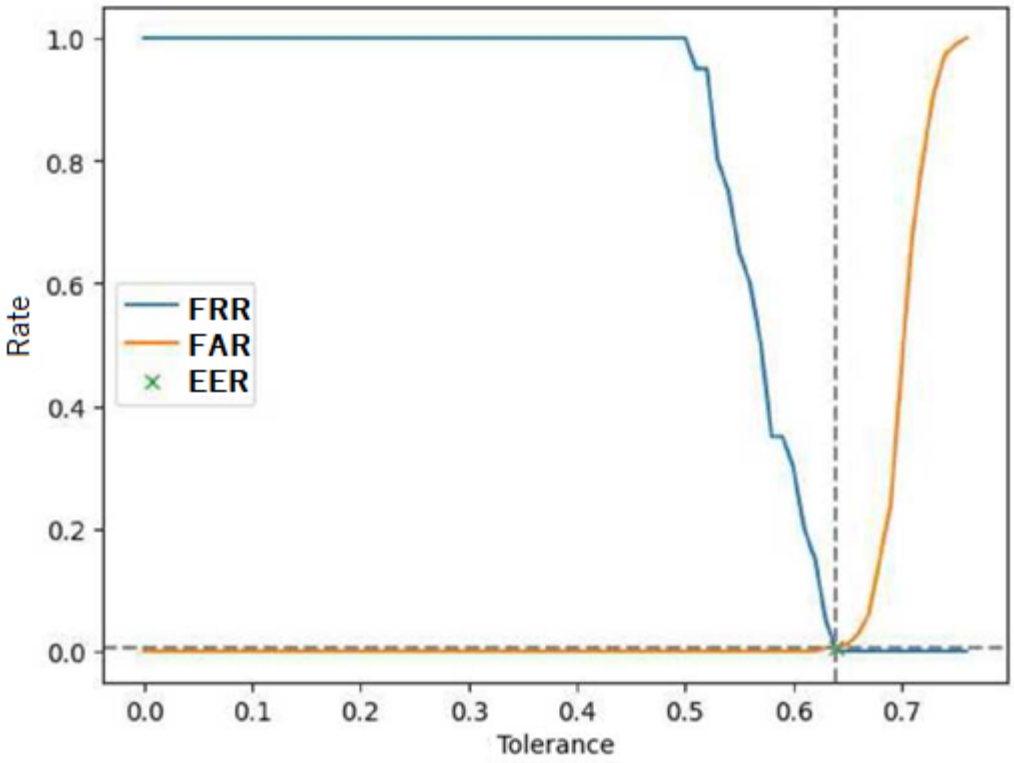

Fig. 4. False Rejection Rate (FRR) and False Acceptance Rate (FAR) curves plotted against system tolerance after fusion.

This figure shows the False Rejection Rate (FRR) and False Acceptance Rate (FAR) curves as a function of system tolerance, after fusion of facial features and skin color with Manhattan distance. The evolution of the two curves is the same as that of the curves in fig. 4, with the particularity that they meet at the point (0.6389; 0.0055).

The ROC curve gives us:

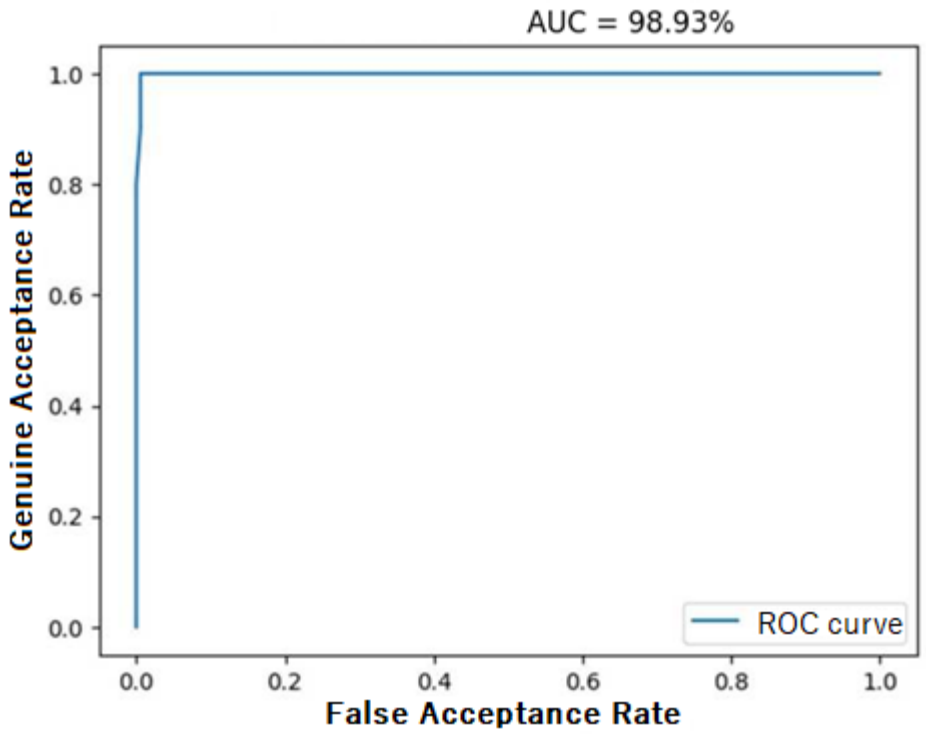

Fig. 5. ROC curve of the system after feature fusion using the Manhattan distance method

The ROC curve shows that after adaptive feature merging, the AUC rises from 98.9% to 98.925%.

These results show that the adapted fusion strategy of face features and skin color with Manhattan distance improves the performance of our system obtained with face only.

3.3. Discussion

In our work, we first implemented a mono-biometric face-based system. Then, we integrated the skin color metadata to validate our initial hypothesis: an adaptive fusion of facial features and skin color would considerably improve recognition performance. The results obtained confirm this hypothesis: we went from a EER of 0.6% to 0.55% and an AUC of 98.9% to 98.925%. Our fusion strategy takes into account the skin color metadata only when it is likely to improve the score. In so doing, our system is sometimes mono-biometric, sometimes multi-biometric. This is worrying in the sense that mono-biometric systems are more prone to problems of performance limitation and vulnerability to identity theft than multi-biometric systems. In the case of impostor images that manage to satisfy the skin color condition before the implementation of our system, we feared a degradation in the robustness of the system obtained after this adaptive fusion compared with the monomodal system designed before the fusion. However, our concerns were allayed after implementation of the adaptive fusion, with the decrease in EER. In fact, EER being the rate of equal error, the point of intersection between False Rejection Rate and False Acceptance Rate is a parameter that perfectly reflects the performance and robustness of a system. So lowering the EER ensures that our adaptive fusion does not increase performance at the expense of robustness, but rather effectively reconciles the two.

4. CONCLUSION

The near-absence of multi-origin biometric systems with a fusion strategy in feature space prompted us to conduct research in this direction. Based on existing and well-known methods such as the HOG descriptor, the FaceNet convolutional neural network, the k-means algorithm and image segmentation in different color spaces, we proposed an adapted fusion of facial features and skin color. The results showed that the joint use of face and skin color recognition led to a significant improvement in the performance of the recognition system compared to using each source of information independently. This strategy improved the performance of the face-only system.

REFERENCES

[1] H. Guesmi, “Identification of people by fusion of differentbiometricmodalities”, EuropeanUniversity of Brittany. Accessed: July 9, 2023. [Online]. Availableat: https://core.ac.uk/reader/ 46813548. French.

[2] A. Jain, P. Flynn, and A. Ross, Handbook of Biometrics. 2008. doi: 10.1007/978-0-387-71041-9.

[3] L. Nouar, “Biometric Identification by Multimodal Fusion”, Thesis, DJILLALI LIABES UNIVERSITY OF SIDI BEL ABBES, Algeria, 2018. Accessed: July 12, 2023. [Online]. Availableat: http://rdoc.univ-sba.dz:8080/jspui/handle/123456789/2418. French.

[4] L. Allano, “Multimodal biometrics: score fusion strategies and dependencemeasuresapplied to virtualpersondatabases”, Doctoral thesis, Evry, National Institute of Telecommunications, 2009. Accessed: March 19, 2023. [Online]. Availableat: https://www.theses.fr/2009TELE0002. . French.

[5] M. Lemmouchi, “Biometric Recognition by Multimodal Fusion”, doctoral dissertation, University of Batna 2, 2020. Accessed:October 26, 2022. [Online]. Availableat: http://eprints.univ-batna2.dz/1866/. French.

[6] Q. N. Nguyen, N. C. Debnath, and V. D. Nguyen, “Face Recognition Based on Deep Learning and HSV ColorSpace,” in Proceedings of the 8th International Conferenceon Advanced Intelligent Systems and Informatics 2022, A. E. Hassanien, V. Snasel, M. Tang, T.-W. Sung, and K.-C. Chang, eds, in Lecture Notes on Data Engineering and Communications Technologies. Cham: Springer International Publishing, 2023, pp. 171-177. doi:10.1007/978-3-031-20601-6_15.

[7] A.-A. Sobabe, T. Djara, B. Blochaou, and A. Vianou, "Soft BiometricsAuthentication: A Cluster-Based Skin Color Classification System", J. Inf. Technol. Res. flight. 15, pp. 1-17, Jan. 2022, doi: 10.4018/JITR.298620.

[8] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, vol. 1. 2005, p. 893. doi: 10.1109/CVPR.2005.177.