Facial Recognition Using Multispectral Images

Luis Carlos Lopes Chambino
Academia Militar, Lisboa
Instituto Superior Técnico (IST), Universidade de Lisboa
luis.chambino@tecnico.ulisboa.pt

Abstract - Facial recognition is a method of identifying or authenticating the identity of people through their faces. Nowadays, facial recognition systems that use multispectral images achieve better results compared to those that use only visible spectral band images. In this work, a skin detector is proposed to be applied in a forgery detection module and an architecture that uses multiple deep convolutional neural networks and multispectral images to perform facial recognition. A study is carried out with the objective of evaluating the performance of the adaptation of several layers of the neural network base. Additionally, a second study was conducted to evaluate the performance of Support Vector Machines (SVM) and k-Nearest Neighbour classifiers to classify the embeddings obtained through the proposed architecture. The experimental results in the Tufts and CASIA NIR-VIS 2.0 multispectral databases show a competitive performance in facial recognition obtaining a rank-1 score of 99.7% and 99.8% respectively.

Keywords - facial recognition, multispectral images, infrared, presentation attack detector.

I. INTRODUCTION

Nowadays, many biometric facial recognition systems work in the visible spectral band. When compared to the various types of biometric traits that exist, such as iris, fingerprint, vein signature and voice recognition, the advantage of using facial recognition over these lies in the possibility of more easily detecting a person's characteristics. Additionally, facial recognition systems are a method whose application is not invasive [1] [2].

Systems that use only the visible spectral band have several obstacles, such as occlusions, variation of poses, cooperation of the person and, the most problematic, changes in luminosity. As a result, it is necessary to complement current facial recognition systems with the use of other biometric sensors (e.g., fingerprint or iris) or with other spectral bands in order to minimize these problems [3].

The use of the infrared electromagnetic spectrum, namely the Near Infrared (NIR), Short Wavelength Infrared (SWIR), Mid Wavelength Infrared (MWIR) and Long Wavelength Infrared (LWIR) spectral bands, has been successfully applied in facial recognition systems, as a complement to the visible spectrum [1] [3]. These systems, which use more than one spectral band, are called multispectral.

Table 1 indicates the most used spectral bands applied in multispectral facial recognition.

TABLE I
SPECTRAL RANGES [4] USED IN FACIAL RECOGNITION.

Spectral Band Name	Wavelength (μm)
Visible	0.38 – 0.75
Near Infrared (NIR)	0.75 – 1.40
Short Wavelength Infrared (SWIR)	1.40 – 3.00
Mid Wavelength Infrared (MWIR)	3.00 – 8.00
Long Wavelength Infrared (LWIR)	8.00 – 15.00

The use of the infrared electromagnetic spectrum in facial recognition systems has several advantages when compared to the electromagnetic spectrum of the visible. Infrared is imperceptible to the human eye and, at the same time, less sensitive to differences in luminosity. For example, the night cameras used in video surveillance have LEDs, with emission in the infrared spectrum in order to illuminate the place and perform night surveillance without people having knowledge [5].

Since NIR and SWIR spectral bands are close to the visible spectral band, it is possible to adapt the trained automatic learning methods with images of the visible spectrum. MWIR and LWIR (also known as thermal) spectral bands allow the use of facial recognition systems at night, when the luminosity is reduced or even zero.

Multispectral facial recognition systems, compared to facial recognition systems, which only use the spectral band of the visible, allow to add a higher level of security and guarantee, for example, in the access to a high security place that the access is made only by authorized people, due to the facial recognition has a higher precision. These places can be hospitals, schools, laboratories, and military buildings [3].

By developing better facial recognition systems, it is possible to ensure more reliable and robust access control, thus protecting property and increasing people's security.

This work is organized as follows: the state of the art study on multispectral facial recognition methods, the most used metrics and public multispectral databases is carried out, in section II. In section III the methodology for multispectral facial detection is defined and proposed. The multispectral databases used are presented in section IV, as well as the results obtained and their respective analysis and discussion.

Section V has the conclusions based on the results and discussion presented in the previous section.

II. STATE OF THE ART

This section summarizes a systematic review of articles in multispectral facial recognition, as well as an analysis of its distribution by years and areas of research, carried out in June of 2020 with the aid of the Web of Science database. Were selected all articles published during the period of January of 2000 to June of 2020, in journals with impact factor (works published in conferences were not considered).

This search located 283 articles published in 132 scientific journals with impact factor. Only articles that perform facial recognition or facial detection with two or more spectral bands (e.g., VIS-NIR, VIS-LWIR, VIS-NIR-LWIR, NIR-LWIR, among other possible combinations) were considered, reducing the number of articles to 47; these papers were considered the most relevant to this work.

An analysis of these articles was carried out taking into consideration the multispectral databases and evaluation metrics used. It was concluded that the most commonly used database was the CASIA NIR-VIS 2.0 [6], used 15 times in the 47 surveyed papers [7].

The metric most used to compare results between methodologies, when using the same database, was the rank-1. This refers to the percentage of predicted identities that return their matching as correct (correctly predicted the person identity), as the highest scoring result (the 1st result).

Through the systematic analysis, the most relevant papers were grouped into five main methods: feature representation, coupled subspace learning, image synthesis, fusion, and deep neural networks. The most used method was deep neural networks, used by 32% of the articles analysed.

The feature representation methods seek to extract the characteristics that are more invariant to the spectral band used. Through the extraction of facial features (e.g., contours, corners, eyes, mouth, among others) it is possible to reduce the information provided by the initial image.

Methods that project the features of different spectral bands into a common subspace are known as coupled subspace learning methods. This subspace allows the identification of the information that is common to the different spectral bands used.

Image synthesis methods transform an image from one spectral band to another spectral band. These methods allow synthesizing an image in the visible spectral band using an image from another spectral band (e.g., LWIR) as a starting point.

The overall performance of the multispectral facial recognition system can be improved by combining several images into a single image, depending on the images used. The most relevant image fusion methods applied in facial recognition are: feature fusion and score fusion; they can be used individually or combined. The feature fusion combines the features of several images, acquired during the feature extraction phase, into a feature vector. Score fusion improves the overall rating performance by combining the output of multiple classifiers into a single classifier.

Currently, the neural network most used in facial recognition is the deep convolutional neural network (DCNN), which comprises a high number of layers when compared to traditional neural networks. DCNN are composed of several layers of convolution, activation, and pooling. The repetition of these layers allows the identification of the most particular and unique features of the images along the neural network. These unique features are denominated embeddings.

Figure 1 shows the distribution of facial recognition methods by year of publication. From the analysis of this graphic, it is possible to conclude the predominance of articles that use the method of image fusion in multispectral facial recognition until 2016 included. From 2017 onwards, most of the papers used deep neural networks, once it provides better results.

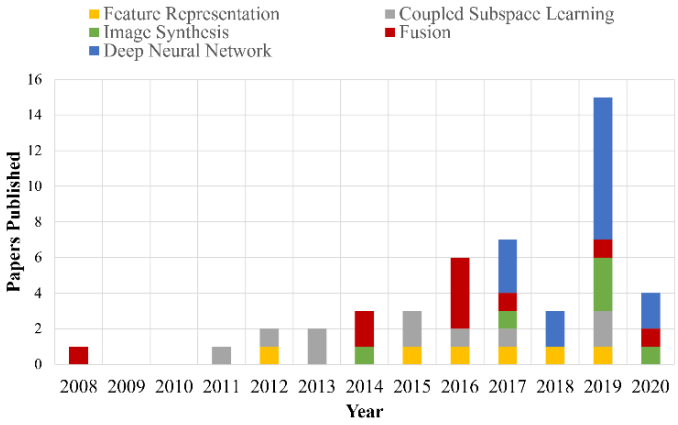

Fig. 1. Distribution of used methods by year of publication.

In Figure 2 it is plotted a boxplot diagram in which it shows the performance obtained in each method. This way is possible to provide a performance comparison of each method.

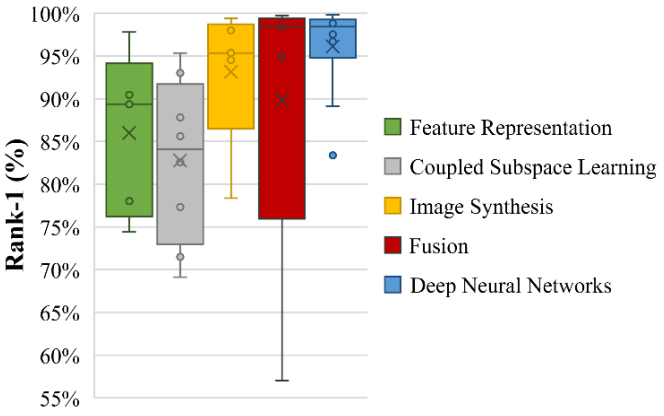

Fig. 2. Comparison of performance by each method.

As it is possible to see from Figure 2, deep neural network obtains the best results, thus justifying the appearance of new neural networks and methods within this area (also proven by Figure 1, year column 2019).

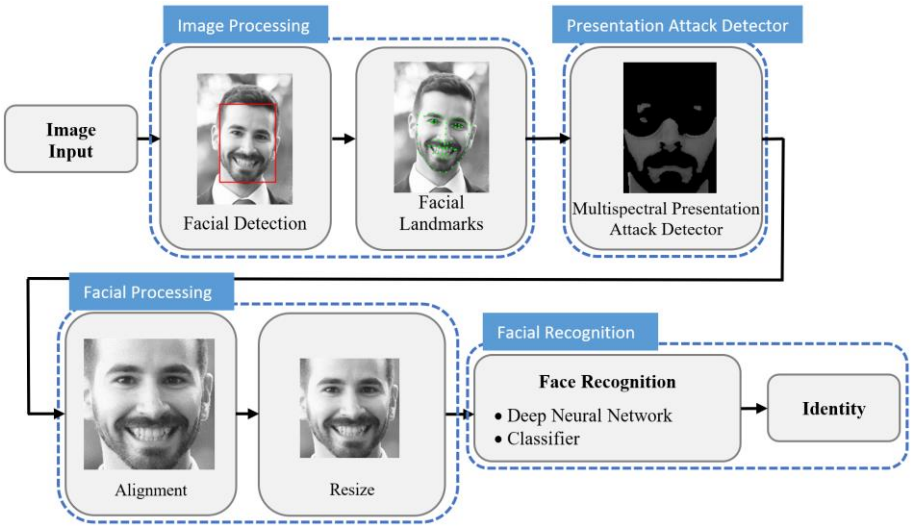

Fig. 3. Flowchart of the proposed multispectral facial recognition system.

III. METHODOLOGY

This section presents the methodology adopted for the implementation of a multispectral facial recognition system. Figure 3 shows the flowchart of the proposed methodology.

The proposed methodology begins with the acquisition of multispectral images (e.g., visible and infrared). These images can be obtained by several monospectral cameras or through an imaging equipment capable of obtaining images at various spectral intervals. The only requirement at this stage is that the images are obtained at the same instant, so that the images have the same person in the same condition of luminosity and pose.

The next step is to convert the images to greyscale, to detect the human faces, and to extract the facial landmarks (e.g., eyes, nose, and mouth) on the image. The module in charge of performing this task is the image processing module.

The proposed facial recognition system includes a module for detecting and warning potential presentation attacks, called the presentation attack detector module. This module takes advantage of all available multispectral images to perform skin detection, thus preventing the facial recognition system from possible presentation attacks.

In the next module facial processing is performed, where the facial landmarks obtained in the first module (image processing) are used in order to align the face. The main objective of this module is to normalize the image before introducing to the DCNN. This facial processing is distinguished from image processing by the fact that the processing is carried out only on the detected face, and not on the global image, as occurs in the image processing module.

A. *Facial Recognition*

The purpose of this module is to extract embeddings representative of the person to be identified through the DCNN, and then a classification of the identity of the person through these embeddings.

In order to extract the embeddings of a facial image, a neural network with an innovative architecture is proposed, shown in Figure 4. Through this neural network it is possible to use several channels, allocating to each channel a spectral band, or spectral range (if several spectral ranges are being used in the same band).

The DCNN used in each channel is the LightCNN [8]. This DCNN stands out from other similar DCNNs because it employs Max-Feature Map (MFM), an extension of the Maxout activation function, in its base architecture. Through this activation function, LightCNN obtains a reduced number of parameters, as an alternative to the rectified linear unit (ReLU) activation function. This network takes as input greyscale images, with a size of 128x128 pixels, and as output, embeddings, representative of the identity of the person, of 256 dimensions.

Different layers in the LightCNN [8] are adapted in order to adjust the model used to a different spectral band, other than the visible spectral band. The channel assigned to the visible spectral band is not modified (in order not to cause overfitting in the data). By reusing the weights of a pre-trained DCNN for facial recognition in a database with a high number of facial images, a possible over-adjustment is avoided, given the limited number of multispectral images used in the training phase [9].

Taking the LightCNN as a starting point, a new layer was added at the end of the network, the final connected layer (FCL), having as an input embeddings with dimensions of Nx256. Through a linear transformation these 256xN embeddings produces the final 256-d embeddings, later used to recognize the face introduced in the neural networks.

Figure 4 shows a generic case of using the proposed network that uses N channels, with the layers that are adapted being marked in green, and the layers that are not adapted in blue. Channel 0, assigned to the visible spectral band, is not adapted.

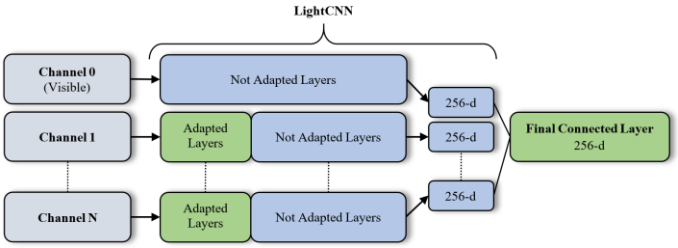

Fig. 4. DCNN proposed architecture scheme.

After obtaining the 256-d embeddings it is necessary to classify them in order to obtain the corresponding identity of the person in the facial image. Several classifiers were tested in order to find out which is the most suitable to classify the 256-d embeddings extracted by the proposed network. The most frequently used classifiers were tested: the SVM with linear or radial basis function (RBF) kernel, and the kNN [10].

Figure 5 shows a summary scheme that exemplifies the facial recognition module, from the input of multispectral images, for each channel, to the identification of the identity of the person present in these images.

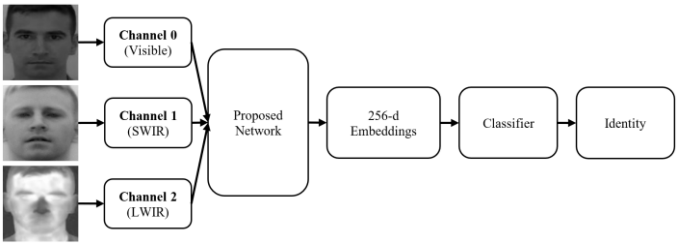

Fig. 5. Diagram of the facial recognition module.

IV. Results and Discussion

This section describes the multispectral databases used, the tests carried out in order to assess which layers of the LightCNN should be adapted, and which classifiers are better to classify a person's identity through the 256-d embeddings. At the end of this section the best classifier is compared with the state of the art methods in multispectral face recognition.

A. Multispectral Databases

In order to correctly evaluate the algorithms included in the proposed methodology, three multispectral databases were used: Tufts [11], CASIA NIR-VIS 2.0 [6] and a set of images acquired at the Portuguese Military Academy (AM).

Before the multispectral databases were used, a cleaning and pre-processing had to be done. The cleaning of the database allowed the exclusion of unusable images (e.g., corrupt or blurred). Then a pre-processing of images was carried out, consisting on the detection and facial alignment in the images, finalized by a cut and resizing of the images present in the multispectral databases. Where it was not possible to make an automatic facial detection, it was necessary to make a manual facial detection.

The Tufts database [11] is composed of three spectral bands, VIS NIR and LWIR. After cleaning, it had a total of 7675 facial images of 109 people, 53 facial images and 4 people were excluded.

CASIA NIR-VIS 2.0 [6] is composed of two spectral bands, VIS and NIR, with 17 489 facial images of 715 people.

The database made at AM is composed of three spectral bands, VIS, SWIR and LWIR. This database was built to be used during the test phase of the presentation attack detector module. During the construction of this database several masks were used, in order to show several presentation attacks.

Figure 6 represents some of the images included in the multispectral database made at the Portuguese Military Academy.

Fig. 6. Example of images from the multispectral database made at the Portuguese Military Academy.

B. Presentation Attack Detector

The tests performed on this module aim to prove the advantage of using multispectral images in presentation attack detection.

A skin detector has been used in the attack presentation detection module, which takes advantage of all available spectral bands in order to perform the skin detection. Next, a comparison is made between what was detected as skin and the facial landmarks, extracted in the image processing module. If the number of facial landmarks that were considered skin is less than 75%, then a presentation attack was detected. In order to prove the efficiency of the proposed multispectral skin detection, a comparison was made with two other skin detectors, the YCbCr skin detector and the HSV.

For this module to be employed, the images must have been acquired in a short time and with a similar frame (i.e., size and pose of the face must be similar in both images at the time of acquisition).

1. Skin Detector

The skin detector performs a pixel-level detection. In the first step, the normalized difference is computed, $d[g_a, g_b]$, for all possible combinations of facial images, taking into account the available channels, using the expression:

$$d[g_a, g_b] = \left(\frac{g_a - g_b}{g_a + g_b}\right) \quad (1)$$

where g corresponds to the pixel intensity value for channel a and b, with $1 \leq a \leq n$ and $a \leq b \leq n$, where n corresponds to the number of channels available in the presentation attack detector module. The normalized difference results in values of $-1 \leq d[g_a, g_b] \leq +1$. Once the normalized difference was computed, it is possible to apply the skin detector that uses the normalized difference values in order to classify the pixels as "skin" or "not skin".

The range of values chosen to classify as "skin" or "not skin", was defined empirically using images in the VIS, SWIR and LWIR spectral bands. The values for skin, when using these spectral bands, are between: $(76, 51, 65) < (d[g_1, g_2], d[g_1, g_3], d[g_2, g_3]) < (131, 140, 127)$.

After skin classification for each normalized difference, a decision is made at the pixel-level. If a pixel is considered "skin" on all normalized difference images, then it is considered as such. A binary map is produced to store all this information, where "1" equals to "skin", and "0" equals to "not skin". This binary map is used later to compute the number of facial landmarks that are considered "skin", for the presentation attack detector.

In Figure 7 the image one the left corresponds to the image before applying the proposed skin detector. The image on the right corresponds to the binary map after applying the proposed skin detector; the black region corresponds to what was classified as "not skin".

Fig. 7. Original image (left) and binary map (map), after applying the proposed skin detector.

To validate our proposal, two other classifiers were used, YCbCr and HSV. Figure 8 show the result obtained with the YCbCr and HSV skin detectors.

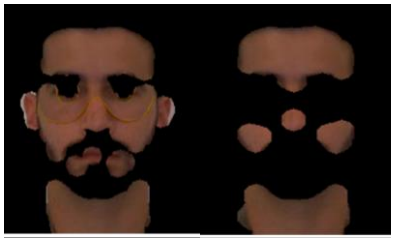

Fig. 8. Results obtained with YCbCr (left) and HSV classifiers, (right).

In a first analysis of Figures 7 and 8, it is possible to observe that the multispectral skin detector is able to make a better skin detection, when compared to YCbCr and HSV skin detectors. The multispectral detector was the only one that could distinguish the real skin from the fake mask skin.

2. Attack Detector

At this stage, the facial landmarks extracted from the previous module (image processing module) are used together with the binary map to detect the presence of a presentation attack.

For this detection, it is computed the percentage of facial landmarks that are considered "skin". If the percentage of facial landmarks is less than 75%, then our multispectral facial recognition system is facing a presentation attack.

The results achieved suggest that the presentation attack detectors that used only the visible spectral band to perform skin detection, have a worst performance than those that used all the spectral bands available. With only the visible spectral band it was achieved a presentation attack detection rate (i.e., when it correctly detected a presentation attack) of 13%, justified by the fact that the used classifiers were not able to make a correct discrimination of the human skin from the mask.

When compared with the multispectral presentation attack detectors, that uses VIS, SWIR and LWIR spectral bands, better results were achieved, the presentation attack detection rate was 83%. Compared with the previous classifiers, it is able to make a correct discrimination of the mask, as shown in Figure 7. Through the achieved results allows to conclude that the excessive luminosity in the images of the spectral band of the visible influence, negatively, the final score of the skin detection.

C. Facial Recognition

After processing the facial images from the databases (i.e., facial detection, alignment, crop and resize), it was necessary to extract the 256-d embeddings from these same images in order to classify the identity of the person present in the images.

Several tests were carried out to find out which layers are the most suitable for neuronal network training; which values should be taken in the hyperparameters of the classifiers; and, after choosing the best hyperparameters, determine the best classifier to classify the 256-d embeddings.

The images of each database were divided into three sets: training, validation, and testing. The percentage of images for the training set was 64%, for the validation set 16%, and finally 20% for the test set. Was performed a stratified division in the database so that each person has an equitable number of facial images of themselves in each set.

1. Training Procedure

Data augmentation was used to obtain a more generalized model. In the images present in the training set,

horizontal random mirroring and a random cropping were used to resize the image to a resolution of 128×128 pixels (the images at the beginning of the network training had a resolution of 144×144 pixels). For the validation set, a crop to the centre was performed, to meet the LightCNN [8] resolution requirement.

During the DCNN training the Cross Entropy (CE) was employed as a loss function. As the DCNN was implemented in Pytorch, the loss function of Cross Entropy combines SoftMax logarithmic (LogSoftMax) and negative log likehood (NLLLoss) in a single loss function.

The batch size was selected so that the numbers of images per batch were as large as possible, to avoid the graphic processing unit (GPU) run out of memory during the training phase. However, it was necessary to ensure that the number chosen for the batch of images was an exponent of 2[1], as suggested by Mishkin [12] and Goodfellow [13].

Table II summarizes the used parameters, for each multispectral database, during the DCNN training procedure of the proposed architecture.

TABLE II
PARAMETERS USED IN THE TRAINING PROCEDURE FOR EACH MULTISPECTRAL DATABASE.

Parameters	Database	
	Tufts	CASIA NIR-VIS 2.0
Batch Size	16	32
Optimization Algorithm	Adam	Adam
Learning Rate	0.001	0.001
Epoch Number	10	50

2. Adapted Layers

Several tests were performed to estimate which layers are suitable to adapt in the LightCNN architecture [8].

Initially, it was only adapted the FCL. This layer initially did not exist in the initial DCNN but was implemented later so that the output of the architecture continued to be the 256-d embeddings. Then, the initial layers of the DCNN were adapted (including the FCL) until all layers from the LightCNN were adapted. In all experiments the weights were initialized from the initial LightCNN model[2] [8].

The nomenclature of the adapted groups followed the initial nomenclature of the LightCNN [8]. LightCNN is composed by 29 layers. In these, 9 sets of layers stand out: the first convolutional layer together with the first MFM, denominated *Conv1*, 4 sets denominated of *Group*, which constitute the layers between the pooling layers, and the remaining 4 layers denominated *Block*, consisting of a block of convolutional layers at the beginning of each Group. The notations used in the combination of the adapted layers are the following:

- **FCL:** Only final connected layer is adapted;
- **Conv1-FCL ({1-1}+FCL):** The first convolutional layer is adapted in conjunction with MFM and FCL;
- **Conv1-Block1- FCL ({1-2}+FCL):** Block of residual neural networks is adapted together with the previous layers;
- **Conv1-Block1-Group1- FCL ({1-3}+FCL):** Adapts Group-1 together with the previous layers;
- **Conv1-N- FCL ({1-N}+FCL):** Adapts layers 1 to N together with the FCL;
- **All Layers:** All layers of LightCNN and FCL are adapted.

The number of epochs used during the training procedure was 10 and 50 for the Tufts [11] and CASIA NIR-VIS 2.0 [6] databases, respectively. After the training, the 256-d embeddings were extracted from each image in the multispectral database. To evaluate each model the SVM-Linear classifier was used to classify the 256-d embeddings.

After the tests were carried out, it was possible to see that, as more layers were adapted, the performance started to deteriorate. The best results were achieved only when the initial layers are adapted. Independently of the database used, for the rank-1 metrics the values of 99.7% and 99.8% were obtained for the Tufts [11] and CASIA NIR-VIS 2.0 [6] multispectral databases, respectively, for the set of layers ({1-3} + FCL).

3. Hyperparameter Analysis

With the best model, obtained in the previous section, the 256-d embeddings were extracted. To classify these embeddings the SVM classifiers (with a linear and RBF kernel) and kNN were used.

In order to make a correct choice in the hyperparameters to be used in each classifier, was used stratified cross validation (SCV), allowing a more correct choice of hyperparameters for unbalanced databases (i.e., number of images per person is not constant in the database), as described by Forman [14] and Tsamardinos [15]. During the SCV the training and validation data set were unified. However, during the training phase of the classifier (with the best hyperparameters already determined) only the training set (i.e., without the validation set) was used.

The use of the SCV is limited by the person who contains in the training and validation set the smallest number of images. The maximum number of times it can be done is 5 times for the Tufts database [11] and 4 times for the CASIA NIR-VIS 2.0 database [6].

The hyperparameter tuned for the SVM-Linear classifier was the regularization parameter (*C*). This hyperparameter indicates the degree of importance given to incorrect classifications. The range of values studied for the *C* hyperparameter was $10^{-10} \leq C \leq 10^{+5}$, with a logarithmic decade interval.

[1] Note that, this limitation is due to the alignment of the virtual processors in the physical processors of the GPU.

[2] The model used is available, for Pytorch, in the following Github repository: https://github.com/AlfredXiangWu/LightCNN.

TABLE III
OPTIMAL VALUES FOR EACH HYPERPARAMETER AND ITS MEAN RANK-1 AND STANDARD DEVIATION USING THE TUFTS DATABASE.

Classifier	Regularization Parameter (C)	Kernel Coefficient (Υ)	Number of Neighbours (k)	Rank-1 (Mean Value)	Rank-1 (Standard Deviation)
({1-3} + FCL) + SVM-Linear	10^{-2}	-	-	99.89 %	0.09 %
({1-3} + FCL) + SVM-RBF	10^{+1}	10^{-4}	-	99.89 %	0.09 %
({1-3} + FCL) + kNN	-	-	1	99.54 %	0.35 %

TABLE IV
OPTIMAL VALUES FOR EACH HYPERPARAMETER AND ITS MEAN RANK-1 AND STANDARD DEVIATION USING THE CASIA NIR-VIS 2.0 DATABASE.

Classifier	Regularization Parameter (C)	Kernel Coefficient (Υ)	Number of Neighbours (k)	Rank-1 (Mean Value)	Rank-1 (Standard Deviation)
({1-3} + FCL) + SVM-Linear	10^{-3}	-	-	99.86 %	0.06 %
({1-3} + FCL) + SVM-RBF	10^{+1}	10^{-5}	-	99.86 %	0.06 %
({1-3} + FCL) + kNN	-	-	1	99.63 %	0.25 %

For the SVM-RBF classifier the following hyperparameters have been refined: the smoothing parameter (C) and the kernel coefficient (Υ). The kernel coefficient hyperparameter aims at defining the influence of a point, in the data set, over others. The range of values studied for the C hyperparameter was $10^{-4} \leq C \leq 10^{+7}$, and to Υ was $10^{-10} \leq \Upsilon \leq 10^{+2}$, both with a logarithmic decade interval.

Finally, for the kNN classifier, the hyperparameter to be tuned was the number of close neighbours (k). This hyperparameter influences the number of points to consider when classifying. The range of values analysed for the k hyperparameter was $1 \leq k \leq 25$.

From the analysis of Tables III and IV, it can be observed that SVM classifiers, regardless of the kernel used, obtain a higher rank-1 score of 99.89% and 99.86% for Tufts [11] and CASIA NIR-VIS 2.0 [6] databases, respectively. These table also show that the best value for the number of neighbours, kNN hyperparameter, is the same, independently of the multispectral database used.

4. Comparison with State of the Art Methods

To identify the most appropriate hyperparameters for each classifier, it is necessary to evaluate the performance of each classifier for the test set of each multispectral database. Simultaneously, an analysis for different classifications is performed using a cumulative correspondence characteristic curve (CMC) to assist in the decision of the classifier. This curve traces the identification rate on the ordinate axis and the rank-N on the abscissa axis.

Using the values in rank-1 and the values obtained for different classifications (i.e., CMC curve) it is possible to determine the best classifier for each multispectral database.

Results using the Tufts database

After evaluating the test set, from the Tufts database, with the classifiers (i) SVM-Linear, (ii) SVM-RBF and (iii) kNN, the following values were obtained in rank-1: (i) 99.7%, (ii) 99.2% and (iii) 99.2%.

Figure 9 shows the CMC curve, for the three classifiers, for the first ten classifications (i.e., rank-10) for the Tufts database [11].

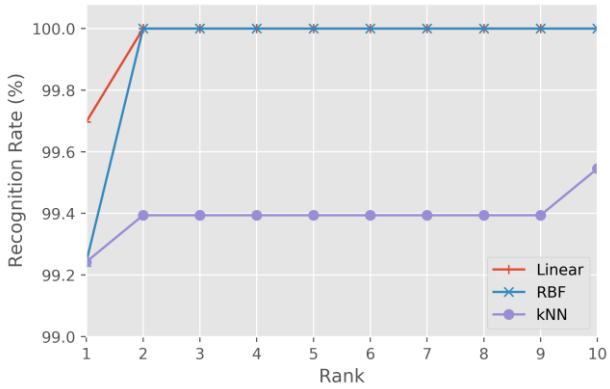

Fig. 9. CMC curve for the SVM-Linear, SVM-RBF and kNN classifiers, for the Tufts multispectral database test set.

Performing a comparative analysis between the three classifiers through the rank-1 values, the SVM-Linear classifier is the one that obtains the best results, with a rank-1 score of 99.7%. Comparatively, the SVM-RBF and kNN classifiers both scored 99.2% for the same set of images.

Figure 9 also shows that for rank-2 and regardless of the kernel used in the SVM classifier, it gets an identification rate of 100%. That is, all the facial images in the test set were correctly identified. On the other hand, the kNN classifier only achieves an identification rate of 100% in rank-102.

Table V presents the results produced by the proposed methodology and by other methodologies described in the literature. In bold is highlighted the method that produced the best score in rank-1.

The proposed methodology uses the LightCNN as base DCNN, adapting the layers ({1-3} + FCL), to produce a set of 256-d embeddings, which are later classified by the SVM-Linear classifier.

TABLE V

RESULTS PRODUCED THROUGH THE PROPOSED METHODOLOGY WHEN COMPARED WITH THE STATE OF THE ART FOR THE TUFTS DATABASE.

Method	Rank-1	Year of Publication
TR-GAN [16]	88.7 %	2019
Circular HOG [17]	94.5 %	2020
Proposed Methodology[3]	**99.7 %**	**2020**

Table V shows that the proposed methodology produces a very competitive result compared to the results produced by other methodologies. When counting the 26 excluded images[4], a rank-1 score of 95.9% is obtained. However, this result is still higher than the state of the art for this database.

It should be noted that as Tufts database is recent, available to the public for research in 2020, the number of researchers using this database is still small.

Results using the CASIA NIR-VIS 2.0 database

After processing the CASIA NIR-VIS 2.0 database test set, with the classifiers (i) SVM-Linear, (ii) SVM-RBF and (iii) kNN, the following values were obtained in rank-1 of (i) 99.8%, (ii) 99.8% and (iii) 99.7%.

Figure 10 shows the CMC curve, for the three classifiers, the first ten classifications for the CASIA NIR-VIS 2.0 database.

In a comparative analysis between the three classifiers through the rank-1 values, both SVM classifiers, regardless of the kernel used, obtain a score of 99.8%. In comparison, the kNN classifier achieves a rank-1 score of 99.7%.

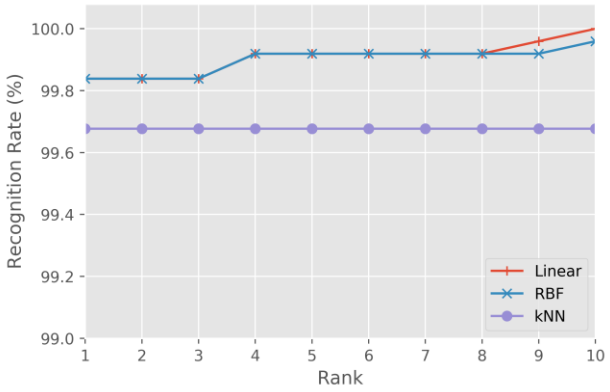

Fig. 10. CMC curve for the SVM-Linear, SVM-RBF and kNN classifiers, for the CASIA NIR-VIS 2.0 test set.

Figure 10 shows that from rank-8 the SVM-Linear classifier achieves a higher identification rate when compared to the SVM-RBF. SVM-Linear and SVM-RBF classifiers get an identification rate of 100% for rank-10 and rank-12, respectively.

[3] It should be noted that the Tufts multispectral database used for DCNN training was cleaned by us. Other authors do not specify whether, or not, they have cleaned the database.

[4] Note that 53 facial images were excluded from the Tufts multispectral database. However, 27 facial images were excluded because 2 people did not had images in all spectral bands, namely VIS, NIR and LWIR.

Table VI shows the results obtained using the proposed methodology and using other methodologies described in the literature. The method that obtained the best score in rank-1 is highlighted in bold. Note that the table is listed by year of publication and not the rank-1 obtained by the methods.

The proposed methodology uses the LightCNN as base DCNN, adapting the layers ({1-3} + FCL), to produce a set of 256-d embeddings, which are later classified by the SVM-Linear classifier.

TABLE VI

RESULTS OBTAINED THROUGH THE PROPOSED METHODOLOGY WHEN COMPARED WITH THE STATE OF THE ART FOR THE CASIA NIR-VIS 2.0 DATABASE.

Method	Rank-1	Year of Publication
CDFL [18]	71.5 %	2015
MCA [19]	69.1 %	2016
MTC-ELM [20]	89.1 %	2017
CEFDA [21]	85.6 %	2017
Oh *et al.* [22]	97.5 %	2017
LightCNN [8]	96.7 %	2018
MDNDC [23]	98.9 %	2019
Peng *et al.* [24]	96.7 %	2019
DSU [9]	96.3 %	2019
WCNN [25]	98.7 %	2019
DDFLJM [26]	98.8 %	2019
Peng *et al.* [27]	98.7 %	2019
CFC [28]	98.6 %	2019
CycleGAN [29]	99.4 %	2020
Proposed Methodology	**99.8 %**	**2020**

From Table VI can be seen that the proposed methodology obtains superior results in rank-1 when compared to the values of other methods described in the state of the art. The most recent work that uses the CASIA NIR-VIS 2.0 database is the article by Bae *et al.* [29] that obtained a rank-1 score of 99.4%, lower than the result obtained by the proposed methodology, of 99.8%.

After a detailed analysis of Table VI, it is possible to see that LightCNN [8] base methodology obtained a rank-1 score of 96.7%. Through the proposed methodology, it was possible to significantly improve the rank-1 score by 3.1%.

V. CONCLUSION

Multispectral facial recognition systems are still complex and demanding, given the numerous factors to consider at the time of facial detection, extraction of facial landmarks and facial recognition. The main applications of multispectral facial recognition systems continue to be security and surveillance, especially in critical locations such as airports or military classified areas.

In this work, a multispectral facial recognition system has been proposed. This system takes advantage of multispectral images in order to obtain better facial

recognition results. The system is composed of four modules: image processing, presentation attack detector, facial processing and facial recognition.

Additionally, in this study a presentation attack detector is proposed in order to detect the presence of presentation attacks. Our module uses a skin detector to create a binary map. With this map, a comparison is made with the facial landmarks to obtain the percentage of facial landmarks that are skin. If the percentage of facial landmarks that are skin is bellow 75%, then we are in the presence of a presentation attack.

It is proposed a multispectral skin detector, that uses all the available spectral bands. YCbCr and HSV skin detectors are used to compare with our multispectral skin detector. During the test phase the VIS, SWIR and LWIR spectral bands are used. The multispectral presentation attack detector achieves better results than those that use only visible images, achieving a presentation attack detection rate of 83%, compared with 13%.

A new architecture for facial recognition using multispectral images is proposed. This architecture has several channels, in which each one is assigned a spectral band or spectral range. Each channel uses the deep convolutional neural network, LightCNN [8], in order to extract the 256-dimension embeddings. Several layers from the LightCNN are adapted in order to adapt each channel to a specific spectral band. In this process, the channel that will receive images in the spectral band of the visible is not considered. In order to maintain the 256-d embeddings as an output of the architecture, a final connected layer (FCL) was implemented. The purpose of this final layer is to carry out a linear transformation in the totality of the 256-d embeddings into a unique 256-d embeddings.

Several layers of LightCNN have been adapted in order to find out which ones present the best results. Through experimentation it is possible to state that the best layers to adapt, regardless of the multispectral database used, are the initial layers, namely ({1-3} + FCL). This study concluded that the higher the number of layers to be adapted, the worse the final score. The best results occur from the adaptation of the initial layers of the neural network.

To classify the 256-d embeddings extracted several classifiers were tested, the SVM (with linear and RBF kernel), and the kNN. The SVM classifier with linear kernel obtained the best values in rank-1 when compared with the other classifiers, for the two multispectral databases used.

Extensive studies in the multispectral databases demonstrated the superiority of the proposed methodology, and rank-1 values of 99.7% and 99.8% were obtained for the multispectral databases Tufts and CASIA NIR-VIS 2.0. Compared to other methodologies identified in the state of the art, the best scores in rank-1 for these databases was 94.5% and 99.8%, respectively.

ACKNOWLEDGMENTS

This work was supported in part by the Military Academy Research Center (CINAMIL) under project Multi-Spectral Facial Recognition, and by FCT with the LARSyS – FCT Project UIDB/50009/2020.

REFERENCES

[1] A. Jain, A. Ross, and K. Nandakumar, Introduction to Biometrics. Springer, 2011.

[2] R. Munir and R. Khan, “An Extensive Review on Spectral Imaging in Biometric Systems: Challenges and Advancements,” Journal of Visual Communication and Image Representation, vol. 65, no. 1, p. 14–26, 2019.

[3] W. Zhang, X. Zhao, J. Morvan, and L. Chen, “Improving Shadow Suppression for Illumination Ro-bust Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 611–624, 2019.

[4] A. D’Amico, C. Natale, F. Castro, S. Iarossi, A. Catini, and E. Martinelli, “Volatile Compounds Detection by IR Acousto-Optic Detectors,” in Unexploded Ordnance Detection and Mitigation. Springer Netherlands, 2009, pp. 21–59.

[5] S. Hu, N. Short, B. Riggan, M. Chasse, and M. Sarfraz, “Heterogeneous Face Recognition: Recent Advances in Infrared-to-Visible Matching,” in International Conference on Automatic Face Gesture Recognition. Washington DC, USA: IEEE, 2017, pp. 883–890.

[6] S. Li, D. Yi, Z. Lei, and S. Liao, “The CASIA NIR-VIS 2.0 Face Database,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, United States of America: IEEE, 2013, p. 348–353.

[7] L. Chambino, J. Silva and A. Bernardino, "Multispectral Facial Recognition: A Review," in IEEE Access, vol. 8, pp. 207871-207883, 2020.

[8] X. Wu, R. He, Z. Sun, and T. Tan, “A Light CNN for Deep Face Representation With Noisy Labels,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2884–2896, 2018.

[9] T. D. Pereira, A. Anjos, and S. Marcel, “Heterogeneous Face Recognition Using Domain Specific Units,” IEEE Transactions on Information Forensics and Security, vol. 14, no. 7, pp. 1803–1816,2019.

[10] X. Fu, J. Lu, X. Zhang, X. Yang, and I. Unwala, “Intelligent In-Vehicle Safety and Security Monito-ring System with Face Recognition,” in IEEE International Conference on Computational Science and IEEE International Conference on Embedded and Ubiquitous Computing Engineering. New York, United States of America: IEEE, 2019, pp. 225–229.

[11] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, “A Comprehensive Database for Benchmarking Imaging Systems,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 3, pp. 509–520, 2020.

[12] D. Mishkin, N. Sergievskiy, and J. Matas, “Systematic Evaluation of Convolution Neural Network Advances on the Imagenet,” Computer Vision and Image Understanding, vol. 161, no. C, pp. 11–19, 2017.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[14] G. Forman and M. Scholz, “Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement,” ACM SIGKDD Explorations Newsletter, vol. 12, no. 1, p. 49–57, 2010.

[15] I. Tsamardinos, A. Rakhshani, and V. Lagani, “Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization,” International Journal on Artificial Intelligence Tools, vol. 24, no. 5, p. 1540023, 2015.

[16] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian, “TR-GAN: Thermal to RGB Face Synthesis with Generative Adversarial Network for Cross-Modal Face Recognition,” in Mobile Multimedia/Image Processing, vol. 11399. SPIE, 2020.

[17] S. Rajeev, K. Shreyas, Q. Wan, K. Panetta, and S. Agaian, “Illumination Invariant NIR Face Recognition Using Directional Visibility,” Electronic Imaging, Image Processing: Algorithms and Systems, pp. 273–1–273–7, 2019.

[18] Y. Jin, J. Lu, and Q. Ruan, “Coupled Discriminative Feature Learning for Heterogeneous Face Recognition,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 3, pp. 640–652, 2015.

[19] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, “Mutual Component Analysis for Heterogeneous Face Recognition,” ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 3, pp. 1–23,2016.
[20] Y. Jin, J. Li, C. Lang, and Q. Ruan, “Multi-task Clustering ELM for VIS-NIR Cross-Modal Feature Learning,” Multidimensional Systems and Signal Processing, vol. 28, no. 3, pp. 905–920, 2017.
[21] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, “Heterogeneous Face Recognition: A Common Encoding Feature Discriminant Approach,” IEEE Transactions on Image Processing, vol. 26, no. 5, pp. 2079–2089, 2017.
[22] B. Oh, K. Oh, A. Teoh, Z. Lin, and K. Toh, “A Gabor-based Network for Heterogeneous Face Recognition,” Neurocomputing, vol. 261, pp. 253–265, 2017.
[23] W. Hu, H. Hu, and X. Lu, “Heterogeneous Face Recognition Based on Multiple Deep Networks With Scatter Loss and Diversity Combination,” IEEE Access, vol. 7, pp. 75 305–75 317, 2019.
[24] C. Peng, N. Wang, J. Li, and X. Gao, “DLFace: Deep Local Descriptor for Cross-Modality Face Recognition,” Pattern Recognition, vol. 90, pp. 161–171, 2019.
[25] R. He, X. Wu, Z. Sun, and T. Tan, “Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1761–1773, 2019.
[26] W. Hu and H. Hu, “Discriminant Deep Feature Learning Based on Joint Supervision Loss and Multi-layer Feature Fusion For Heterogeneous Face Recognition,” Computer Vision and Image Understanding, vol. 184, pp. 9–21, 2019.
[27] C. Peng, N. Wang, J. Li, and X. Gao, “Re-Ranking High-Dimensional Deep Local Representation for NIR-VIS Face Recognition,” IEEE Transactions on Image Processing, vol. 28, no. 9, pp. 4553–4565, 2019.
[28] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, “Adversarial Cross-Spectral Face Completion for NIR-VIS Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1025–1037, 2019.
[29] H. Bae, T. Jeon, Y. Lee, S. Jang, and S. Lee, “Non-Visual to Visual Translation for Cross-Domain Face Recognition,” IEEE Access, vol. 8, no. 7, pp. 50 452–50 464, 2020.