Multispectral Face Recognition on the Wild

Pedro Martins
Academia Militar, Lisbon, Portugal
IST-UL, Lisbon, Portugal
pedro.roque.martins@tecnico.ulisboa.pt

Alexandre Bernadino
IST-UL, Lisbon, Portugal
ISR, Lisbon, Portugal
alex@isr.tecnico.ulisboa.pt

José Silva
Academia Militar, Lisbon, Portugal
LIBPhys-UC, Coimbra, Portugal
jose.silva@academiamilitar.pt

***Abstract*—This work proposes a multispectral face recognition system in an uncontrolled environment, aiming to identify or authenticate identities (people) through their facial images. Face recognition systems in uncontrolled environments have shown impressive performance improvements over the last decades. However, most are limited to the use of a single spectral band in the visible spectrum. The use of multispectral images makes it possible to collect information that is not obtainable in the visible spectrum when certain occlusions exist (e.g., fog and plastic materials) and in low or no light environments. The proposed work uses the scores obtained by face recognition systems in different spectral bands to make a joint final decision in identification. The evaluation of different methods for each of the components of a face recognition system allowed selecting the most suitable ones for a multispectral face recognition system in an uncontrolled environment. The experimental results, expressed in Rank-1 scores, were 99.5% and 99.6% in the TUFTS multispectral database with pose variation and expression variation, indicating that the use of multispectral images in an uncontrolled environment is advantageous when compared with the use of single spectral band images.**

***Index Terms*—deep neural networks, multispectral face recognition, on the wild, score fusion.**

I. Introduction

MOST of the current face recognition systems require the cooperation of the user to ensure that pictures are taken in favourable conditions (frontal postures, good illumination, no occlusion) and have trouble dealing with uncontrolled scenarios. Uncontrolled environment scenarios, such as riots and violent demonstrations, can often be used by criminals and terrorist cell members to move around and cause damage to National Security, as this type of environment adds difficulty to their detection. The uncontrolled environment is mainly characterized by [1]: variety of lighting; variety of pose; variety of facial expressions and existence of occlusions. These features are challenges to face recognition systems due to the multiple intrapersonal variations they provide, making it difficult to correctly identify an individual's identity based on a collaborative image of the individual.

This work has as its main objective the development of a multispectral face recognition system in an uncontrolled environment. To achieve this goal, the solutions used by current recognition systems and the evaluation of the benefits of using multispectral images are explored. The developed face recognition system is evaluated in public multispectral image datasets with pose and expression variability.

II. Background

A. Face Recognition

In general, a face recognition system is described by several phases. The **first** phase consists of acquiring the facial images and pre-processing them, such as locate the faces and crop them them. In a **second** phase a set of features is extracted from the facial image, for instance the position of facial landmarks, eye distance or even the face tones. Finally, these features are used in a classifier for Identification or Verification purposes.

Face recognition can be performed in a controlled or uncontrolled environment. The controlled environment is one in which the user cooperates in the recognition by facilitating it through correct and static posture in a place with good lighting. In the uncontrolled environment, recognition is dynamic, without the user cooperating in acquiring an image, making the face recognition process very difficult due to the diversity of the surrounding environment and facial poses and expressions.

B. Multispectral Imaging in an Uncontrolled Environment

The databases of the visible (VIS) domain and the use of image synthesizers, that generate multiple poses and facial expressions from the obtained images, have allowed circumventing the difficulties associated with the variety of pose and facial expressions. However, two points have proved more difficult to overcome: the change of illumination and occlusions. This has motivated the use of multiple spectral bands, with particular emphasis on the Infrared (IR) spectral band, that can acquire images in environments with little or no brightness and overcome occlusions such as smoke and fog. In short, multispectral analysis allows a face recognition system to extract facial features that would be impossible to obtain with images from the VIS spectral band. The use of IR images for automatic face recognition is not without challenges, as these images are sensitive to the emotional, physical and health conditions of the individual, as well as the surroundings, and do not serve as an absolute alternative to the use of the VIS spectrum, but rather as a complement [2].

III. Related Work

A. Face Recognition in an Uncontrolled Environment

The uncontrolled environment, strongly characterized by pose-light-expression factors, emerges as a problem for current recognition systems. A significant step was taken towards solving this type of problem by introducing very large databases to train *Deep Convolutional Neural Networks* (DCNN) in combination with the emergence of image synthesis methods [1]. The two main image synthesis methods are: (i) one-to-many augmentation, which consists in generating different poses of a face from a canonical face image; (ii) many-to-one normalization, which consists in normalizing any pose of the face to a canonical face pose [1]. Since the appearance of *Generative Adversarial Networks* (GAN) in face normalization, with DR-GAN [3], GANs have taken the lead in solving the problem of pose and facial expression variation. As for one-to-many augmentation using GANs, as is the case with the DA-GAN network [4], their image production power also gives them an advantage compared to other algorithms.

B. Multispectral Face Recognition

The main multispectral face recognition methods can be categorized in Fusion and Image Synthesis.

Fusion methods are subdivided into feature fusion and score fusion. In the first, a fusion of features from the different spectral bands of the facial image is performed, allowing extracting the most relevant features from the different bands and joining them in a vector. The second method combines the scores obtained from each classifier uni-band *versus* uni-band (e.g. a classifier operating only in the Long-Wavelength IR (LWIR) band and another operating only in the Near IR (NIR) band). Examples of this type of method are those proposed by Seal et al. [5] and Kanmani et al. [6].

The image synthesis methods allow transforming an image of a spectral band into another, helping compare two images. The main advantage of image synthesis is that it enables passing an image from any spectral band to the VIS band, making it possible to use classifiers implemented to process images of the VIS spectrum [7]. One of the most recent works in this area synthesizes VIS images from NIR images using GANs [8].

C. Gaps

Although several papers address multispectral face recognition, few of these demonstrate its power in an uncontrolled environment due to the limitations in current databases of multispectral face images. In existing datasets, the variations of conditions are not extreme, being usually semi-controlled environments and not *on the wild* (uncontrolled environment). Thus, the fact that these databases are incomplete (compared to those of the VIS band) is still a barrier to improving the capability of multispectral face recognition systems in an uncontrolled environment.

The present work proposes a system that integrates the capabilities of current face recognition systems in an uncontrolled environment in the VIS spectrum at the pose variation level and the capabilities of multispectral face recognition systems to surpass illumination variation.

IV. METHODOLOGY

The proposed multispectral face recognition system consists of three tasks: (i) Face Detection and Alignment, (ii) Image Synthesis and (iii) Face Recognition.

In Fig. 1, the general operation of the proposed face recognition system is shown, including the steps performed in each task.

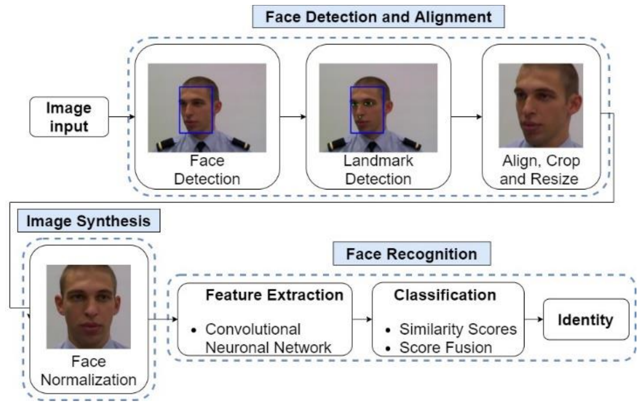

Fig. 1. Schematic of the operation of the proposed face recognition system.

In the initial phase of the system, it is necessary to acquire multispectral images. After image acquisition, the Face Detection and Alignment module aims to obtain an aligned and centred facial image with predefined dimensions. To achieve this goal, it is necessary to detect the presence of human faces and then perform a face marking, detecting essential landmarks of the face, such as eyes and nose, allowing a correct alignment of the face and cropping around it. The following task is Image Synthesis, which aims to obtain a frontal facial image. The next task is Face Recognition, where facial image features are extracted through a CNN and a one-shot learning methodology is followed for the classification task, obtaining similarity scores for each spectral band. These scores are combined using a score fusion method, and the predicted identity is the one with the highest combined score.

A. Face Detection and Alignment

Face detection, in conjunction with face alignment, aims to detect the faces presented in the input image and identify facial landmarks so that faces are centred, aligned, and equally sized. The procedure of face detection and alignment does the following: given an image, identifies the different faces present, extracts the facial landmarks, and processes the image to produce facial images where the face is centred and aligned.

The face detection algorithms explored in this work (S3FD [9], OpenCV [10] and DSFD [11]) are based on SSD (single-shot multibox detector), a deep learning architecture for object detection [12]. The basic idea of the SSD is to generate scores for the presence of each object category in each predefined box and produce adjustments to the box to match the shape of the object. As for the facial landmark detection algorithms, the DLIB [13] and 2D-FAN [14] were tested. Both networks receive an image of a person and produce, as output, the position of the different facial landmarks around the face. All the algorithms were trained in databases in the spectral band of the VIS.

B. Image Synthesis

To overcome the problems associated with image acquisition in an uncontrolled environment, such as variation in lighting, occlusions and changes of poses, a face normalization module is used. This module aims to synthesize (create) an image of a face with frontal pose and neutral expression from a non-frontal face image. The models FNM [15] and FFWM [16] are analyzed. FNM is a GAN that uses a network specialized in obtaining facial features to provide the ability to preserve facial identity and facial discriminators to refine local textures. The normalization method of the FFWM model consists of using a deformation module, aiming to synthesize realistic frontal images with illumination preservation.

C. Face Recognition

This last module aims to identify the person present in an input face image. For this purpose, it is necessary to perform two tasks: feature extraction and classification.

The extraction of representative features from a facial image is performed through a version of Light CNN [17] with 29 convolutional layers (Light CNN-29). To use this network for feature extraction in spectra other than VIS, transfer learning is used. According to [18], several models for biometric recognition are based on transfer learning when the databases are limited. Thus, one should use the Light CNN-29 model with the weights obtained by training on the VIS databases and fine-tune with the facial image databases in spectra other than the VIS. At the end of the feature extraction phase, *B* vectors of 256 dimensions are generated, being *B* the number of spectral bands in which the facial image was acquired.

The classification process applied by the one-shot learning technique determines the degree of similarity of the input image with the images of each class present in the support set, which is constituted by one example per class that the classifier has access to its identity The similarity functions to be used are the Euclidean distance and the cosine similarity. After obtaining the similarity values for each identity in the different spectral bands, a fusion of the obtained scores is performed, inspired by [24]:

$$S_{ic} = \sum_{b=1}^{B} S_{ib} W_b \tag{1}$$

where S_{ic} is the combined score for each identity i and S_{ib} is the score obtained for each band b for each identity i. W_b is the weight of each spectral band. The weights associated with each band are fixed, determined by the accuracy obtained when classifying with only that band [19]. In this way, the band that usually obtains the most reliable similarity scores to classify will have a greater weight in the fusion of scores. The prediction is then made by choosing the identity i of the support set that has the highest combined similarity score:

$$prediction = max(S_{ic})\ \forall i \in [1, \dots, N] \tag{2}$$

V. Results and Discussion

A. Databases

We performed both qualitative and quantitative evaluation of the proposed methods. Qualitative evaluation methods use images obtained from the Military Academy to visualize the behaviour of the different algorithms. These images are in the VIS, NIR and LWIR bands. TUFTS [20] multispectral database was used for quantitative evaluation. The TUFTS database has facial images in the VIS, NIR and LWIR bands of 113 people with different poses and different illumination conditions. The TUFTS database has different subsets, divided into TUFTS-Pose (facial images with 9 different poses per individual, in Visible, NIR and LWIR) and TUFTS-Exp (4 facial images with different expressions and one with sunglasses per individual, in Visible and LWIR) to study pose variation and expression variation separately.

B. Metrics

The metrics used are Rank-1, Rank-5 and TAR@FAR=0.001. When using a generic expression Rank-n, given an image of a face as input, the classifier obtains the *n* most probable identities, one of which is the correct identity. TAR (true accept rate) is defined as the percentage of faces that, compared to the corresponding gallery identity, are identified as matches, while FAR (false accept rate) is the percentage of incorrect identities to which a face is matched.

C. Face Detection and Alignment

1) Face Detection

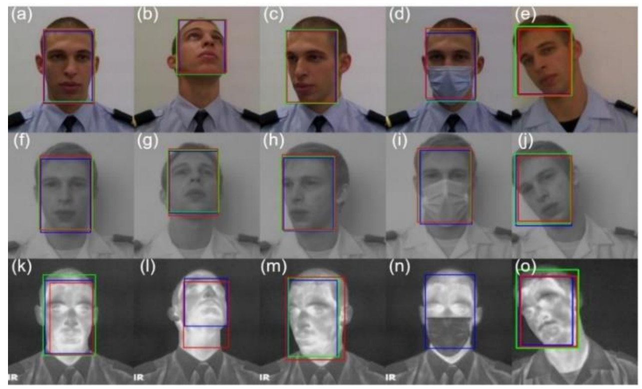

Fig. 2. Results obtained by facial detection methods in the spectral bands of VIS (above), NIR (middle) and LWIR (below). S3FD-red, DSFD-blue, OpenCV-green.

Regarding the qualitative results presented in Fig. 2, all algorithms produced similar results in the VIS band. This was expected since they were all trained in databases of the spectral band of the VIS. In the LWIR spectral band, a failure of the OpenCV network was observed in the second facial pose, where it cannot detect any face. The DSFD maintained the same results in VIS and LWIR bands, being a good indicator of its ability to extract characteristics even in the LWIR spectral band.

TABLE I

Accuracy of the different face detection algorithms in the TUFTS database.

	Accuracy at different spectral bands (%)		
Method	**VIS**	**NIR**	**LWIR**
OpenCV	99.2	90.4	77.7
S3FD	**99.9**	**100.0**	90.8
DSFD	**99.9**	**100.0**	**98.8**

The quantitative results are presented in TABLE I. It can be observed that the OpenCV network results are lower than the others, especially in infrared bands. We observe that the DSFD maintains a very high accuracy for the different spectral bands, thus being the best network for face detection in a multispectral facial analysis system.

2) Landmark Detection and Facial Alignment

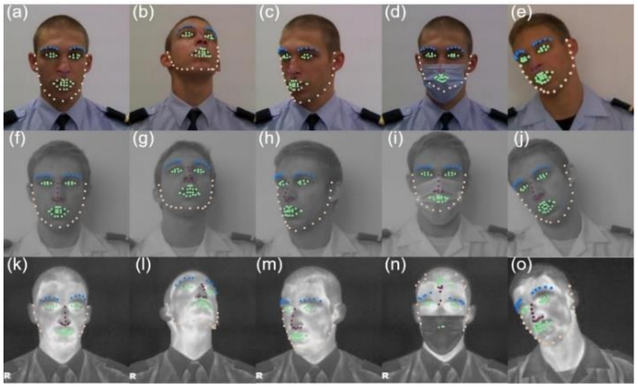

Fig. 3. Results achieved by 2D-FAN in the spectral bands of VIS (above), NIR (middle) and LWIR (below). Yellow-jawline, green-eyes and mouth, purple-nose, blue-eyebrows.

Given the results obtained we decided to use the 2D-FAN (results present in Fig. 3) over the DLIB's network due to two factors: (i) it showed better results with face pose variation and (ii) it is the only one capable of producing positive results in the LWIR spectral band.

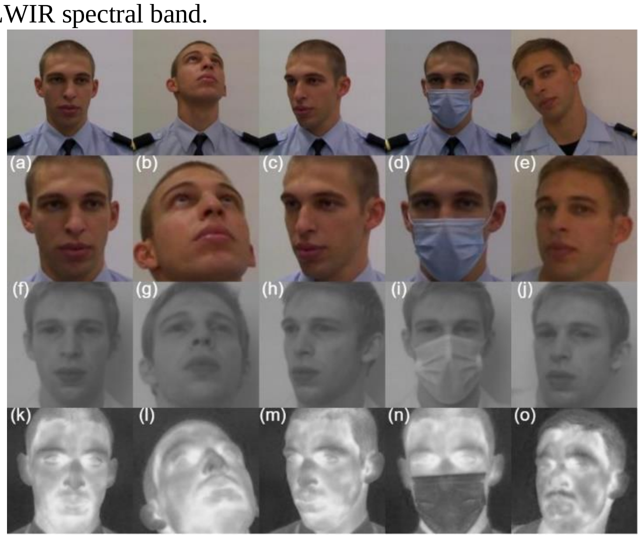

Fig. 4. Results achieved by the proposed facial detection and alignment module in the different spectral bands. The images on the top are the originals in the VIS.

After the face detection with DSFD and landmark face detection with 2D-FAN, the align, crop, and resize phase took place, which aligned the imaginary eye line of all detected faces with the horizontal, centred the faces in the images, cropped them and resized to the same size, resulting in the results presented in Fig. 4. The alignment effect is strongly noticeable on the rightmost facial image. This normalization of the facial images can help a multispectral face recognition system in an uncontrolled, where faces can be presented in several poses.

D. Image Synthesis

For all images used in the qualitative and quantitative evaluations, the images were previously processed to be properly centred, aligned and scaled. The rightmost images used in the previous tasks were replaced by ones with a strong expression, to evaluate the capacity of the models to normalize expressions.

1) Selecting the Best Model

Fig. 5. Results achieved by the FNM in the different spectral bands. The images on the top are the originals in the VIS. The second, third and fourth row are the images produced from the VIS, NIR and LWIR images, respectively.

Given the results obtained, we decided to use the FNM (see Fig. 5) instead of the FFWM due to two factors: (i) the FFWM requires a specific face detection and alignment module and that the face is perpendicular to the horizontal, while the FNM is more robust to pose variations in the input image; (ii) all images normalized by the FNM tend to maintain the face proportions, without deforming them, in the NIR and VIS spectral bands.

It is also relevant to point out that the FNM normalizes pose and expression, eliminates face masks, as is the case of the surgical mask, and normalizes to the VIS spectral band. However, this normalization doesn't produce realistic results with the LWIR images due to the difference between the LWIR and VIS spectral bands.

2) Evaluation of Selected Model

Identification with and without the use of FNM was performed to verify its advantage. For this purpose, the Light CNN-29 was used for feature extraction, and the identification was performed based on the score obtained by cosine similarity.

TABLE II
RESULTS (IN %) WITH AND WITHOUT FNM ON THE TUFTS-POSE DATABASE.

	Rank-1		**Rank-5**		**TAR @FAR=0.001**	
	w/o	**w/**	**w/o**	**w/**	**w/o**	**w/**
VIS	80.3	**96.2**	91.0	**99.5**	60.8	**87.2**
NIR	98.3	**99.0**	99.5	**99.8**	90.4	**91.9**
LWIR	**41.8**	34.9	**58.2**	57.8	**28.7**	14.0

The results presented in TABLE II show that, without using the FNM, the use of the NIR spectral band produces better results than the VIS band in all metrics analyzed. A possible explanation is that the images obtained in the NIR band are not so affected by the illumination variation (due to pose variation), thus not causing as many occlusions as in the VIS band.

The results improve with the use of the FNM in the VIS and NIR spectral bands, with increases in performance in Rank-1 of 15.9% and 0.7%, respectively. In the remaining metrics, it is also observed better values with the use of the normalization model. The results in the LWIR spectral band indicate that using the FNM does not improve the performance in any of the metrics, so the FNM model is used only to normalize facial images from the TUFTS-Pose database in the VIS and NIR spectral bands.

E. Face Recognition

1) Network training

For the training phase, and considering the results presented above, it was decided to make only one fine adjustment to the LWIR band feature extraction network. In order to train the Light CNN-29 with identities (people) different from the test ones, a last connected layer was added for training purposes and the LWIR spectral band images from the IRIS database [21] were used. In this way, Light CNN-29 can be applied to other databases to extract features from facial images to serve as input for similarity functions. Thus, all the following processes make use of the 256-dimensional feature set obtained by Light CNN-29. The values used to the different parameters during the 10 epochs of training were: 16 to batch size, 10^{-4} to Learning Rate and 0.9 to momentum, following the suggested by [17].

The fine-tuning allowed the network to improve the rank-1 in the LWIR band on the TUFTS-Pose from 41.8% to 55.5% and on TUFTS-Exp from 67.5% to 79.6%.

2) Score-level Fusion

At this stage, we have three Light CNN-29 models, each responsible for extracting features from a specific band. Only the Light CNN-29 responsible for the extraction of features from the LWIR spectral band underwent a fine-tuning.

Using the scores obtained by each spectral band, it is now possible to proceed to the final classification. A fusion of the achieved scores was performed using (1). Two studies were conducted, with different weights of each band. In study 1, the previously obtained test results are not considered thus, the same weight is used in all spectral bands. The Wb values in study 2 are derived from the mean of the Rank-1 average precision of each of the spectral bands in the tests performed on the TUFTS-Pose, TUFTS-Exp and CASIA NIR-VIS 2.0 [22] databases with the cosine similarity function, rounded to tenths.

TABLE III
RESULTS (IN %) OBTAINED IN THE FACE RECOGNITION TASK, IN THE TUFTS-POSE DATABASE.

	Rank					TAR @FAR =0.001
	1	2	3	4	5	
Study1	99.4	**99.8**	99.9	**100.0**	**100.0**	90.5
Study2	**99.5**	**99.8**	**100.0**	**100.0**	**100.0**	**93.5**
VIS	96.2	98.7	99.1	99.4	99.5	87.4
NIR	99.0	99.7	99.7	99.8	99.8	93.1
LWIR	55.6	62.2	66.7	69.9	72.6	30.5

TABLE III presents the results obtained with the TUFTS-Pose database. These results show that study 2 achieved better results than study 1, in the Rank-1 and Rank-3 metrics by 0.1 percentage points, and the TAR@FAR=0.001 metric by 3 percentage points. The superiority of the results obtained by study 2 compared to study 1 shows that the weight assigned to the LWIR spectral band should be lower than the weight assigned to the others because the characteristics obtained in the LWIR spectral band are the least representative of the identity.

Analyzing the results of the different spectral bands separately, the NIR spectral band achieved the best results due to its robustness towards the variation of illumination present in the TUFTS-Pose database. Despite the promising results of the NIR band when used solo, study 2 obtained superior results in all metrics, with particular emphasis on Rank-1 (from 99.0% to 99.5%) and TAR@FAR=0.001 (from 93.1% to 93.5%). It is relevant to point out that only the results obtained with score fusion reached the 100% accuracy rate in the assessed Ranks (Rank-4 for study 1 and Rank-3 for study 2).

TABLE IV
RESULTS (IN %) ACHIEVED IN THE FACE RECOGNITION TASK, USING THE TUFTS-EXP DATABASE.

	Rank					TAR @FAR =0,001
	1	2	3	4	5	
Study1	**99.6**	**100.0**	**100.0**	**100.0**	**100.0**	98.7
Study2	**99.6**	**100.0**	**100.0**	**100.0**	**100.0**	99.3
VIS	**99.6**	99.6	99.8	**100.0**	**100.0**	**99.4**
LWIR	79.6	86.3	88.5	90.4	91.6	54.9

TABLE IV shows the results obtained with the TUFTS-Exp database. An analysis of the results allows us to see that the face recognition results obtained are better with score fusion, where both studies obtained the same result as the VIS spectral band in Rank-1 (99.6%) but managed to achieve a higher result in Rank-2 (100% against 99.6% of the VIS spectral band). However, the best result for TAR@FAR=0.001 is obtained using only the VIS spectral band, with 99.4%, while the second-best result was obtained in study 2, with 99.3%.

Performing a global analysis of all results, we can observe that the fusion of scores mainly favours cases where the results obtained by the different spectral bands separately were less satisfactory. Looking at the results obtained with the TUFTS-Exp, it is clear that the VIS spectral band already obtains very high values in all metrics. This fact makes the fusion of scores not so effective. However, despite a decrease of the TAR@FAR=0.001 in TABLE IV, the results obtained by the fusion of scores, in general, were higher than those obtained by the spectral bands separately. The results obtained thus demonstrate the benefit of using multispectral images in a face recognition system.

VI. CONCLUSIONS

In this paper, a multispectral face recognition system in an uncontrolled environment has been proposed, aiming to make a decision with the largest amount of data available, i.e. using the facial images obtained by the different spectral bands. The system is composed of three modules: (i) face detection and alignment, (ii) image synthesis and (iii) face recognition.

Several techniques were implemented to validate them in different multispectral bands, since all of them were trained on Visible databases, as well as to analyze the influence of facial image features (pose, illumination and expression). This analysis aimed to select the most appropriate technique for each module of the proposed face recognition system.

The extraction of the feature sets of the facial images from the different spectral bands is performed using Light CNN-29 [69], with a fine adjustment to the network weights for the LWIR spectral band since it was trained on the Visible spectral band. For the classification phase, identification is performed in the different spectral bands, each producing different scores for each identity. In this work, two different studies were performed for score fusion, which allowed us to conclude that: (i) simply using the different spectral bands to identify is advantageous (study 1) and (ii) a weighted average is beneficial when the different classifiers have different levels of reliability (study 2).

On the multispectral TUFTS database, the results obtained in Rank-1 by the proposed system and with score fusion with a weighted average (study 2) regarding pose variation and expression variation were 99.5% and 99.6% respectively, with the best results obtained using only one spectral band being 99.0% (in NIR) and 99.6% (in VIS), as showed in Table III and IV. On the TAR@FAR=0.001 metric, the results obtained by weighted average (study 2) regarding pose variation and expression variation were 93.5% and 99.3%, while with only one spectral band 93.1% (in NIR) and 99.4% (in VIS) were obtained, showed in the Table III and IV.

As contributions to state of the art, the analysis of several techniques for different tasks stands out. This analysis allowed: (i) to present an efficient face detection and alignment module to be used by any multispectral face analysis system, (ii) to identify the situations in which the FNM model should be used to normalize facial images and (iii) the selection of a similarity function and the weights to be used in the fusion of scores to identify/verify identities. From the experimental results, it is also concluded that the proposed system allows obtaining high results in multispectral face recognition in an uncontrolled environment, where the use of the scores obtained from different spectral bands allows, in general, to achieve superior results than using only the scores obtained by one spectral band.

Acknowledgements

This work was supported in part by the Centro de Investigação da Academia Militar (CINAMIL) under the project Multi-Spectral Facial Recognition, and by FCT with the project LARSyS - FCT UIDB/50009/2020.

References

[1] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, “Deep face recognition: A survey,” in *31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, 2018, pp. 471–478.

[2] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, “Improving shadow suppression for illumination robust face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 611–624, 2018.

[3] L. Tran, X. Yin, and X. Liu, “Disentangled representation learning GAN for pose-invariant face recognition,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1283–1292.

[4] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, “3d-aided dual-agent gans for unconstrained face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, 2018.

[5] A. Seal, D. Bhattacharjee, M. Nasipuri, C. Gonzalo-Martin, and E. Menasalvas, “Fusion of visible and thermal images using a directed search method for face recognition,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 04, p. 1756005, 2017.

[6] M. Kanmani and V. Narasimhan, “Optimal fusion aided face recognition from visible and thermal face images,” *Multimed. Tools Appl.*, pp. 1–25, 2020.

[7] C. Peng, X. Gao, N. Wang, and J. Li, “Graphical representation for heterogeneous face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, 2016.

[8] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, “Adversarial cross-spectral face completion for NIR-VIS face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, 2019.

[9] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3FD: Single Shot Scale-Invariant Face Detector,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 192–201.

[10] B. Gary, “The OpenCV Library,” *Dr. Dobb’s J. Softw. Tools*, vol. 25, no. 2236121, pp. 120–123, 2008.

[11] J. Li *et al.*, “DSFD: Dual shot face detector,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5055–5064.

[12] W. Liu *et al.*, “SSD: Single shot multibox detector,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9905 LNCS, pp. 21–37.

[13] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[14] A. Bulat and G. Tzimiropoulos, “How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks),” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[15] Y. Qian, W. Deng, and J. Hu, “Unsupervised face normalization with extreme pose and expression in the wild,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9843–9850.

[16] Y. Wei, M. Liu, H. Wang, R. Zhu, G. Hu, and W. Zuo, “Learning Flow-Based Feature Warping for Face Frontalization with Illumination Inconsistent Supervision,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12357 LNCS, pp. 558–574, 2020.

[17] X. Wu, R. He, Z. Sun, and T. Tan, “A light cnn for deep face representation with noisy labels,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.

[18] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, “Biometrics Recognition Using Deep Learning: A Survey,” *CoRR*, vol. 1912.00271, 2019.

[19] N. Srinivas, K. Veeramachaneni, and L. A. Osadciw, “Fusing correlated data from multiple classifiers for improved biometric verification,” in *12th International Conference on Information Fusion*, 2009, pp. 1504–1511.

[20] K. Panetta *et al.*, “A Comprehensive Database for Benchmarking Imaging Systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, 2020.

[21] IEEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under grant R01-1344-18; FAA/NSSA grant R01-1344-48/49, Office of Naval Research under grant #N000143010022., “Dataset 02: IRIS Thermal/Visible Face Databases, ” 2005. http://vcipl-okstate.org/pbvs/bench/ (accessed Mar. 27, 2021).

[22] S. Z. Li, D. Yi, Z. Lei, and S. Liao, “The CASIA NIR-VIS 2.0 face database,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.