A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums

Virginia Espinosa-Duro · Marcos Faundez-Zanuy · Jiří Mekyska

Received: 28 December 2011 / Accepted: 27 June 2012 / Published online: 18 July 2012 © Springer Science+Business Media, LLC 2012

Abstract In this paper, we present a new database acquired with three different sensors (visible, near infrared and thermal) under different illumination conditions. This database consists of 41 people acquired in four different acquisition sessions, five images per session and three different illumination conditions. The total amount of pictures is 7,380 pictures. Experimental results consist of single sensor experiments as well as the combination of two and three sensors under different illumination conditions (natural, infrared and artificial illumination). We have found that the three studied spectral bands contribute in a nearly equal proportion to a combined system. Experimental results show a significant improvement combining the three spectrums, even when using a simple classifier and feature extractor. In six of the nine studied scenarios, we obtained identification rates higher or equal to 98 %, when using a trained combination rule, and two cases of nine when using a fixed rule.

Keywords Thermal image · Visible image · Near-infrared image · Face recognition data fusion

Introduction

Face is one of the most challenging traits for biometric recognition [1]. Real-world tests of automated face recognition systems have not yielded encouraging results. For instance, face recognition software at the Palm Beach International Airport, when tested on fifteen volunteers and a database of 250 pictures, had a success rate of less than fifty percent and nearly fifty false alarms per five thousand passengers, which means two to three false alarms per hour per checkpoint [2]. Even when the face recognition task is performed by a human operator, it is far from perfect and errors exist. Humans can recognize faces from different views. However, they do this accurately only when the faces are well known to them.

The limits of human performance do not necessarily define upper bounds on what is achievable. Specialized identification systems, such as those based on novel sensors, may exceed human performance in particular settings [2]. For this reason, it is interesting to perform automatic experiments with images acquired with different sensors. To this aim, we studied in our previous work [3, 4] whether there is complementary information when acquiring an image with sensors that acquire the face in different frequency ranges.

A long-standing focus of research in human perception and memory centers on the importance of the “average” or “prototype” in guiding recognition and categorization of visual stimuli. The theory is that categories of objects, including faces, are organized around a prototype or average. The idea is that the closer an item is to the category prototype, the easier it is to “recognize” as an exemplar of the category. However, in biometric applications, the goal is not to detect a face in an image. We must determine whether the face is known to us and whose face it is [2].
The authors of [5] found that faces rated as “typical” were recognized less accurately than faces rated as “unusual.” In general, artists draw caricatures emphasizing facial features that are “unusual” in average population and are present in the caricaturized person. Despite the fact that caricatures are grotesque distortions of a face, they are often recognized more accurately and efficiently than actual images of the faces [5]. Computer-generated caricatures likewise operate by comparing a face to the “average face” and then by exaggerating facial dimensions that deviate from the average [6]. The enhanced recognizability of caricatures by comparison to veridical faces may be due to the fact that the exaggeration of unusual features in these faces makes the person less confusable with other faces, and somehow or other “more like themselves” [2].

We would expect that the recognizability of individual faces should be predicted by the density of faces in the neighboring “face space.” We might also expect that the face space should be most dense in the center near the average. The space should become progressively less dense as we move away from the average. If a computationally based face space approximates the similarity space humans employ for face processing, we might expect that “typical” faces would be near the center of the space and that unusual or distinctive faces be far from the center. It follows, therefore, that computational models of face recognition will not perform equally well for all faces. These systems should, like humans, make more errors on typical faces than on unusual faces.

On the other hand, computer-based systems can go beyond human limitations because they can “see” beyond cognitive limits. For this purpose, we created a new database. Although several databases exist that simultaneously acquire visible and near-infrared [7] or visible and thermal images [8], we are not aware of an existing database containing visible, near-infrared and thermal information simultaneously.

In this paper, we present a new database and we extract quantitative measurements in three spectral bands: visible, near infrared and thermal.

The paper is organized as follows: section two describes the new database. Section three presents the face segmentation and normalization procedure, feature extraction, classifier and experimental results. Section four summarizes the main conclusions.

### Face database

We have acquired a database of 41 people using three sensors simultaneously. Next sections describe the details of this new database acquired in visible, near-infrared and thermal spectrums.
We have designed a background screen using a special stand kit, which supports a roll of matt black paper. It is important to point out that this matt black background is mandatory behind the user in order to avoid undesirable thermal reflections from the operator, due to its well-known extra low albedo. This smooth background also facilitates the segmentation of the visible and NIR images.

Lighting conditions

In each recording session, the images have been acquired under three different illumination conditions:

(a) Natural illumination (NA): Windows are open and sunlight enters the room. Obviously, this illumination is not constant along days (due to weather conditions) and it also varies in function of the different hours of the day.

(b) Infrared illumination (IR): Printed circuit board around the webcam is turned on and the remaining sources of light are disconnected. A graphic user interface has been developed in order to set properly the IRED’s intensity level and to set the image involved parameters (exposure, gamma and brightness). Additionally, it is also possible to manually fully optimize them.
Artificial illumination (AR): The provided equipment used for illumination is the following: A set of 9 cool white fluorescents uniformly distributed in order to produce the base illumination of the scene. A second pair of IANIRO Lilliput lights fitting 650 W–3,400 K tungsten halogen lamps have also been used in order to fill and smooth the well-known discontinuous fluorescent spectral emission and to provide an additional IR portion of light. Figure 4 shows the related portion of spectral emission in this band emitted by a set of different color temperature halogen bulbs.

At the beginning, high pair of power focus produced important dark shadows over the users’ face. In order to solve this drawback, we had finally used a LEE 3ND 209 Filter to minimize the referred effect. This neutral density gel reduces light without affecting color balance.

Acquisition protocol

Each user has been recorded in four different acquisition sessions performed between November 2009 and January 2010. In this sense, distinctive changes in the haircut and/or facial hair of some subjects may be appreciated. The acquisitions have been done in the whole day from 9 AM to 5 PM, because it was getting dark after 5 PM. The average time required for the full acquisition process of a skilled user has been 10 min, being 15 min for a nonskilled one. The whole set of users were acquired in 2 days per session.

The time slot between each session is shown in Table 1.

In each illumination condition, five different frontal snapshots are acquired. During the acquisition process, the user is required to look straight at the same place. No keeping neutral facial expression is required. Thus, a different facial expressions have been collected (smiling/nonsmiling, open-closed and blinking eyes, etc.). Due to glasses exhibit a fully different behavior as function of the spectrum, being transparent from the VIS to the NIR spectrum and fully opaque beyond 3 μm approximately as showed in Fig. 5, people wearing glasses were asked to remove them before acquisition. No any other physical restriction has been taken into account in order to acquire a face image.

In order to reduce the correlation between consecutive acquisitions of the same session, between a couple of snapshots the user is asked to stand up, make a loop to the room, including one step that corresponds to the portion of the room close to the blackboard, and sit down again. It is worth to mention that thermal camera was able to detect a temperature increase due to this additional physical exercise.

Database features

Final database consists of 41 people (32 males, 9 females). Each individual contributed in four acquisition sessions (see Table 1) and provided five different snapshots in three different illumination conditions and under three image sensors. This implies a total of: $41 \times 4 \times 5 \times 3 \times 3 = 7,380$ images, grouped in folders shown in Fig. 6.

In order to normalize all the images to the same size and remove the background, we have used a Viola and Jones face detector [9]. However, it was unable to segment correctly the thermal images, and a new face segmentation algorithm for thermal images has been developed [10]. All the faces have been segmented and consequently resized to 100 × 145 pixels using bicubic interpolation.

The images in NIR spectrum are stored in lossless *.bmp files. The images from thermal camera were firstly stored to *.bmt format provided by TESTO company. This file includes VIS image, temperature matrix and metadata describing, for example, the outside humidity, temperature range, etc. This file was processed and the image in VIS spectrum was extracted. The temperature matrix was stored to MATLAB *.mat file and also transformed to grayscale image and stored to *.bmp format.

Each file in database has an 8-letter code name. The meaning of each letter is described in Table 2.
**Experimental results**

In order to compare the identification rates using different sensors and illumination conditions, we have used a simple feature extraction method based on discrete cosine transform (DCT). According to our previous experiments [11], this method outperforms the well-known eigenfaces [12] algorithm with lesser computation burden.
Feature extraction algorithm

Given a face image, the first step is to perform a two-dimensional DCT, which provides an image of the same size but with most of the energy compacted in the low frequency bands (upper left corner).

The discrete cosine transform (DCT) is an invertible linear transform and is similar to the discrete Fourier transform (DFT). The original signal is converted to the frequency domain by applying the cosine function for different frequencies. After the original signal has been transformed, its DCT coefficients reflect the importance of the frequencies that are present in it. The very first coefficient refers to the original signal’s lowest frequency and usually carries the majority of that is not so important. The goal is to reduce the dimensionality of the vectors in order to simplify the complexity of the classifier and to improve recognition accuracy. In this paper, we will follow a frequency selection mechanism by means of discriminability criteria. The goal is to pick up those frequencies that yield a low intra-class variation and high inter-class variation. On the other hand, those frequencies that provide a high variance for inter- and intra-class distributions should be discarded. The notation is the following one:

- $P$ is the number of images per person in the training subset.
- $F$ is the number of people inside the database.
- $i_{pf}(x,y)$ is the luminance of a face image $f$ that belongs to person $p$, where $p = 1, \ldots P; f = 1, \ldots F$.
- $I_{pf}(f_1, f_2) = \text{transform}\{i_{pf}(x,y)\}$ is the DCT2 transformed image.
- $m(f_1, f_2) = \frac{1}{PF} \sum_{p=1}^{P} \sum_{f=1}^{F} I_{pf}(f_1, f_2)$ is the average of each frequency obtained from the whole training subset images.
- $m_p(f_1, f_2) = \frac{1}{F} \sum_{f=1}^{F} I_{pf}(f_1, f_2) \quad \forall p = 1, \ldots P$ is the average of each frequency for each person $p$.
- $\sigma_p^2(f_1, f_2) = \frac{1}{F} \sum_{f=1}^{F} (I_{pf}(f_1, f_2) - m_p(f_1, f_2))^2, \forall p = 1, \ldots P$ is the variance of each frequency for each person $p$.
- $\sigma^2(f_1, f_2) = \frac{1}{PF} \sum_{p=1}^{P} \sum_{f=1}^{F} (I_{pf}(f_1, f_2) - m(f_1, f_2))^2, \quad \forall p = 1, \ldots P$ is the variance of each frequency evaluated over the whole training subset.
- $\sigma_{\text{intra}}^2(f_1, f_2) = \sum_{p=1}^{P} \sigma_p^2(f_1, f_2)$ is the average of the variance of each frequency for each person.

We can define a zonal mask as the matrix $m(f_1, f_2) = \begin{cases} 1, & f_1, f_2 \in L \vspace{0.5em} \ \text{and multiply the transformed image by} \\ 0, & \text{otherwise} \end{cases}$

the zonal mask, which takes the unity value in the zone to be retained and zero on the zone to be discarded. In image coding, it is usual to define the zonal mask taking into account the transformed coefficients with largest variances. In image coding, the goal is to reduce the amount of bits without appreciably sacrificing the quality of the reconstructed image, and in image recognition, the number of bits is not so important. The goal is to reduce the dimensionality of the signal with the higher frequencies. These coefficients generally represent greater image detail or fine image information, and are usually noisier. DCT has an advantage when compared with DFT: the coefficients are real values while DFT produces complex values. We used DCT2 (two-dimensional DCT) defined by the following equations:

Figure 7 summarizes the process to obtain a feature vector from a DCT transformed image.

Feature extraction using DCT consists of selecting the coefficients around the $X[0,0]$ coefficient (DC coefficient), where the highest discrimination capability between different people is.

We can define a zonal mask as the matrix $m(f_1, f_2) = \begin{cases} 1, & f_1, f_2 \in L \vspace{0.5em} \ \text{and multiply the transformed image by} \\ 0, & \text{otherwise} \end{cases}$

the zonal mask, which takes the unity value in the zone to be retained and zero on the zone to be discarded. In image coding, it is usual to define the zonal mask taking into account the transformed coefficients with largest variances. In image coding, the goal is to reduce the amount of bits without appreciably sacrificing the quality of the reconstructed image, and in image recognition, the number of bits is not so important. The goal is to reduce the dimensionality of the signal with the higher frequencies. These coefficients generally represent greater image detail or fine image information, and are usually noisier. DCT has an advantage when compared with DFT: the coefficients are real values while DFT produces complex values. We used DCT2 (two-dimensional DCT) defined by the following equations:

$$X[k, l] = \frac{2}{N} c_k c_l \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} x[m, n] \cos \left( \frac{(2m + 1)k\pi}{2N} \right)$$

$$\times \cos \left( \frac{(2n + 1)l\pi}{2N} \right) \quad (1)$$

where in Eq. (1):

$$c_k, c_l = \begin{cases} \sqrt{\frac{1}{2}} \quad \text{to } k = 0, l = 0 \\
1 \quad \text{to } k = 1, 2, \ldots a - 1 \text{ and } l = 1, 2, \ldots b - 1 \end{cases} \quad (2)$$

The process to obtain a two-dimensional DCT plus a feature selection.
• \( \sigma_{\text{inter}}^2(f_1, f_2) = \sigma^2(f_1, f_2) \).

We will use the following measure, which is the Fisher discriminant:

\[
M_1(f_1, f_2) = \frac{m_{\text{intra}}(f_1, f_2) - m_{\text{inter}}(f_1, f_2)}{\sigma_{\text{intra}}^2(f_1, f_2) + \sigma_{\text{inter}}^2(f_1, f_2)} \tag{3}
\]

It is interesting to point out that this procedure is similar to the threshold coding used in transform image coding [13]. Nevertheless, we are using a discriminability criteria, instead of a representability criteria, which is only based on energy (the higher the frequency coefficient value, the higher its importance).

Figure 8 shows an example of the \( M_1 \) ratio obtained from visible images of session 1.

It is important to point out that feature selection has been done using only training samples. Thus, we have not selected frequencies using testing samples, which would provide better results, but unrealistic because feature selection must be done a priori, before classifying the samples.

Classification algorithm

We will use a simple distance calculation between training and testing feature vectors of dimension \( N \) using a fractional distance. We have also successfully applied this classifier in [14] for signature recognition and in [15] for speaker recognition. It is represented in Eq. (4):

\[
d(x, y) = \left( \sum_{i=1}^{N} (|x_i - y_i|)^{p/2} \right)^{1/p} \tag{4}
\]

where \( i \) is the feature vector component.

For \( p = 2 \), the equation corresponds to the Euclidean distance. When data are high dimensional, however, the euclidean distances and other Minkowsky norms (\( p \)-norm with \( p \) being an integer number, that is, \( p = 1; 2; \ldots \)) seem to concentrate, and so, all the distances between pairs of data elements seem to be very similar [16] Therefore, the relevance of those distances have been questioned in the past, and fractional \( p \)-norms (Minkowski-like norms with an exponent \( p \) less than one) were introduced to fight the concentration phenomenon. In our case, we have used \( p = 0.5 \).

We have experimentally selected the number of coefficients (vector dimension) by trial and error, selecting a window of \( 1 \times 1, 2 \times 2, 3 \times 3, \ldots N \times N \), where the frequency coefficients have been previously ordered using the strategy defined in Sect. 3.1.

We have used a simple method because the experiments are quite time-consuming. For each feature vector dimension we have executed the algorithm, we have studied hundreds of feature vector dimensions for each condition, and this implies thousands of executions. If using a more sophisticated method, this would imply, probably, to train a complex algorithm for each studied feature vector dimension. This would be impractical from the computational burden point of view. In fact, in its current version, we required several weeks to work out the whole experimental section.

In addition, we were looking for a method with few parameters because a more complex algorithm can require fine tuning, and this fine tuning could be different for each spectral band. Thus, in this case, it would be difficult to know whether one spectral band provides better results due to different tuning or to the frequency itself. Our suggested method is so simple and effective that we did not required any fine tuning.

Experimental results with different illuminations

In this section, we compare the identification rates for the visible (VIS), near-infrared (NIR) and thermal (TH) sensors for natural (NA, Fig. 9), artificial (AR, Fig. 10) and infrared (IR, Fig. 11) light. These experimental results have been obtained training with sessions 1 and 2 and testing with session 3 as function of \( N \) (see Fig. 7). Thus, the number of selected coefficients for each point in these plots is \( N^2 \).

These figures reveal several general interesting facts:

• Feature selection is indeed important, because a too large number of coefficients decrease the identification rate.

• Different sensors provide a different number of optimal feature dimension \( N^2 \).

Figure 9 shows the following:

• The NIR sensor provides lower identification rates than visible and thermal, which provide similar rates. In
addition, the optimal feature vector size is more critical, because identification rates drop quickly when moving away from the optimal point.

- The TH sensor requires a lesser amount of coefficients to reach the highest identification rate, and the identification rate drops slower than for visible sensor.

Figure 10 shows the following:

- All the sensors provide nearly similar results, although the visible sensor outperforms the other ones.
- Optimal feature vector size selection is lesser critical for the VIS sensor than for the other ones because a large range of $N^2$ values produce the highest achievable identification rate.

Figure 11 shows the following:

- The NIR sensor provides the best behavior, and the VIS sensor fails to provide a reasonable identification rate. This makes sense considering that infrared illumination in the proposed scenario for a visible sensor is equivalent to an almost dark scene.
- TH and NIR provide similar behavior, although TH sensor results drop faster beyond the optimal value.

Experimental results for a specific sensor

In this section, we compare the identification rates for a specific sensor regarding the different illuminations. We
have studied the VIS sensor (Fig. 12), the NIR (Fig. 13) and the TH (Fig. 14) for natural (NA), artificial (AR) and infrared (IR) illumination.

Figure 12 reveals the following:
- VIS sensor performs better with artificial illumination. This makes sense because the variation along acquisition sessions is smaller than when using natural light, which varies from day to day.
- Optimal feature selection value is more critical when using natural light when compared to artificial light.
- VIS sensor fails when using NIR illumination. This is due to the acquisition conditions for this scenario, which is almost dark for a visible sensor.

Figure 13 reveals the following:
- IR sensor performs similarly well with AR and IR illumination, and around 10% worse when evaluated with natural light. This can be due to the larger variability when analyzing faces with natural light.
- Feature selection is less critical when using IR illumination. This is reasonable considering that NIR sensors should perform optimally with IR illumination.

Figure 14 shows an expected conclusion:
- TH sensor performs almost the same with all the studied illuminations. This is reasonable considering that thermal cameras do not measure the light reflection.
Fig. 13 Identification rate as function of the square size \((N)\) of selected coefficients for Infrared sensor and natural (NA), artificial (AR) and near-infrared (NIR) illumination.

![Identification rate plot for IR sensor](image)

Fig. 14 Identification rate as function of the square size \((N)\) of selected coefficients for thermal sensor and natural (NA), artificial (AR) and near-infrared (NIR) illumination.

![Identification rate plot for TH sensor](image)

Table 3 Optimal results for visible (VIS), near-infrared (NIR) and thermal (TH) sensor under natural (NA), infrared (IR) and artificial (AR) illumination conditions.

| Sensor | Illumination | NA Identification | NA Coefficients | IR Identification | IR Coefficients | AR Identification | AR Coefficients |
|--------|--------------|-------------------|----------------|-------------------|----------------|-------------------|----------------|
| VIS    | 89.76        | 20 x 20           |                | 46.83             | 19 x 19        | 92.20             | 18 x 18        |
| NIR    | 80.49        | 16 x 16           |                | 90.73             | 15 x 15        | 91.22             | 11 x 11        |
| TH     | 88.29        | 16 x 16           |                | 90.73             | 14 x 14        | 88.29             | 11 x 11        |

Experimental conditions are the same of previous Figs. 9, 10, 11, 12, 13, and 14. The selected number of coefficients is also represented.
on the face. They measure the heat emission of the body. In fact, they could perfectly work in full darkness because the illumination is irrelevant.

It is important to point out that although there are small variations between the three studied illumination conditions, they are not due to the illumination. The motivation is the inherent variability of the acquired subject from day to day and acquisition to acquisition. If the subject would be an inanimate object with a fix temperature along the different acquisitions, the behavior shown in Fig. 11 would be the same under the three illuminations. However, a human being cannot fulfill this property.

Table 3 summarizes the optimal results and the optimal feature vector dimension (when evaluated from $1 \times 1$, $2 \times 2$, … $N \times N$) for different sensors and illumination conditions. This table reveals similar identification rates for all the sensors, although the thermal one requires a lower number of coefficients. In addition, the visible sensor provides low identification rates when using IR illumination for the reasons previously commented.

Experimental results in mismatch conditions

Using the setup of previous sections, we have studied the identification rates in function of the different illumination conditions for training and testing. Table 4 shows the experimental results when using $20 \times 20$ coefficients for VIS, NIR and TH, respectively. The models have been computed using sessions 1 and 2 and the testing has been done with sessions 3 and 4 separately.

Although it is possible to trade-off an optimal feature vector dimension for each scenario, we decided to select a fix window size of $20 \times 20$ coefficients. According to the previous plots (Figs. 9, 10, 11, 12, 13, 14), this tends to benefit the identification rates of the VIS sensor. Nevertheless, the goal of this table is to study the mismatch illumination effect between training and testing conditions, rather than to find the highest identification rate for each scenario.

Due to the bad results obtained specially when using IR illumination, we have decided to use some normalization procedure. The image has been normalized previous to DCT2. The normalization maps the values in intensity image to new values such that 1% of data is saturated at low and high intensities of the image. This increases the contrast of the normalized image. Thus, Table 4 includes experimental results with and without normalization.

Face recognition in new spectral bands implies new problems that must be addressed. Nevertheless, the solution should be specific to each spectral band. For instance, while general temperature can rise in thermal images, the relative difference between different portions of the face can remain similar, because the hottest point will always be related to the vein positions, and these remain the same. On the other hand, we trained with sessions 1 and 2 and tested...
with sessions 3 and 4. Thus, the experimental results are affected by the time evolution. Nevertheless, we have applied a feature selection algorithm that looks for low intra-class variation and high inter-class variation. Thus, stability along time is achieved by means of feature selection (see Sect. 3.1), which is different for each spectral band. Comparing the experimental results of testing sessions 3 and 4, we can speculate that the stability of the different frequency bands over long periods of time is reasonably good, because there is a minor degradation when comparing session 4 and 3.

Table 4 reveals the following aspects:

- IR sensor provides the best result, which is 94.1% identification rate. This experimental result is in agreement with our previous paper’s conclusion, because NIR images have higher entropy than the other ones.
- Looking at the standard deviation (std) and mean value (m) of the experimental results of Table 4 for a specific sensor, we obtain the following values: m = 81.6 and std = 12.4 for visible sensor, m = 69.6 and std = 22.8 for near-infrared sensor and m = 80.9 and std = 3.2 for thermal sensor. Thus, thermal image recognition rates are more stable than the other sensors.
- Image normalization is important for the case of illumination mismatch when using the visible and near-infrared sensor, and less important for the thermal one.

**Experimental results using multi-sensor score fusion**

In this section, we combine the scores provided by different sensors in order to improve the recognition accuracies. The existing fusion levels [17] are sensor, feature, score and decision. Some papers use image fusion and then perform the recognition over this fused image [18–22]. This is known as “sensor fusion.” Another possibility is decision fusion [23]. In our paper, we will use a score combination. Some papers have also studied this possibility [24–27]. In fact, Buyssens and coworkers [25] studied the sensor, feature and score level and found that data fusion at score level outperforms the other ones when combining visible and thermal images. However, Raghavendra and coworkers [26] studied the fusion of visible and near-infrared images and found slightly better accuracies when fusing images than applying other fusion levels. To the best of our knowledge, there is no paper devoted to visible, near-infrared and thermal images using a multisession database. The fusion scheme is presented in Fig. 15.

This kind of fusion is also known as confidence or opinion level. It consists of the combination of the scores provided by each matcher. The matcher just provides a distance measure or a similarity measure between the input features and the models stored on the database.

Before opinion fusion, normalization must be done when the scores provided by different classifiers do not lie in the same range. In our case, we experimentally found that this normalization is not necessary because the three studied classifiers gave similar range.

After the normalization procedure, several combination schemes can be applied [17]. The combination strategies can be classified into three main groups:

a) Fixed rules: All the classifiers have the same relevance. An example is the sum of the outputs of the classifiers. That is, let o1 and o2 be the outputs of classifier numbers 1 and 2, respectively. For example, a fixed combination rule yields the combined output O = (o1 + o2)/2

b) Trained rules: Some classifiers should have more relevance on the final result. This is achieved by means of some weighting factors that are computed using a training sequence. For instance, O = ω1o1 + ω2o2 = ω1o1 + (1 − ω1)o2

c) Adaptive rules: The relevance of each classifier depends on the instant time. This is interesting for variable environments. That is, O = ω1(t)o1 + (1 − ω1(t))o2. For instance, a system that detects a low illumination scene can weight more the thermal score.

The most popular combination scheme is the weighted sum: Oj = \sum_{i=1}^{N} ωio_{ij}

where the weights can be fixed, trained or adaptive.

In this paper, we will use a fixed rule scheme as well as a trained rule, although in our case the purpose of the trained rule is to evaluate the weights assigned to each classifier, rather than to maximize the identification rate. In fact, trained rules should be done with a development set different than the test set. Otherwise, the experimental results are unrealistic. Nevertheless, these optimistic results are well accepted by the scientific community. The same situation occurs, for instance, in biometric verification performance measured by means of equal error rate (EER). EER corresponds to a posteriori threshold setup, which is an optimistic (unrealistic) situation. In “real systems,” the threshold must be setup a priori and then the errors should be computed with testing samples not used for threshold computation. This implies some degradation on experimental results, but it is well accepted this unrealistic (optimistic) situation.

Table 5 shows the identification rates under different training and testing conditions for a fixed rule using the same weight for all the classifiers.
When combining two classifiers using a trained rule, a trial and error procedure must be done to set up the optimal value of the weighting factor. Figure 16 shows the identification rates as function of the weighting factor $\alpha$, where the combination function is as follows:

$$d = \alpha \times d_{\text{VIS}} + (1 - \alpha) \times d_{\text{NIR}}$$

It is interesting to point out that for $\alpha = 1$, the combination consists of the visible classifier distance alone $d_{\text{VIS}}$, while $\alpha = 0$ fully removes the effect of the visible classifier, being the classification based on near-infrared sensor distance alone $d_{\text{NIR}}$. Thus, for $\alpha = 0$ we obtain 89.8% identification rate and for $\alpha = 1$ 84.4%. In the middle, there is an area that provides higher recognition rates (up to 95.6%) due to the combination of distances.

When combining three classifiers, we can generalize the previous procedure using the following combination function:

$$d = \alpha \times d_{\text{VIS}} + \beta \times d_{\text{NIR}} + (1 - \alpha - \beta) \times d_{\text{TH}}$$

In this case, we should trade-off two parameters and the graphical representation is a three-dimensional plot, such as the one shown in Fig. 17. This three-dimensional plot is not very informative due to the limitations of three-dimensional representations and an alternative is to represent its contour plot. A contour plot are the level curves of the bidimensional matrix formed by giving values to the two parameters $\alpha$ and $\beta$. For the sake of simplicity, only a few level curves are plot, as well as a black dot that indicates the highest value.

Some interesting remarks about this kind of plot are the following:

- In fact, the addition of the three weighting factors should be one. However, in order to avoid discontinuities and sudden gradients, we have filled up a whole matrix with $\alpha, \beta \in [0, 1]$ using increments of 0.01. Thus, 100 values have been worked out for each variable.
- $\alpha = 100$ implies $\beta = 0$. Thus, the combined system consists of the visible sensor alone.
- $\beta = 100$ implies $\alpha = 0$. Thus, the combined system consists of the near-infrared sensor alone.
- $\alpha = \beta = 0$ implies that the combined system consists of the thermal sensor alone.
- $\alpha = \beta = 33$ implies that the three systems are equally weighted in the averaged distance computation.
- $\alpha$ and $\beta$ adjustments on the diagonal line depicted in each of the Figs. 18 and 19 imply that the thermal sensor is not used. The closest is the optimal point to this line, the lesser the weight of the thermal system. Adjustment points far from this diagonal imply a strong weight on the thermal system.

Observing the 18 plots of Figs. 18 and 19, it can be said that the three systems are almost equally important in the weighting process. There is only one exception, which is the second plot of Fig. 18. In this case, $\alpha = 33$,
### Table 5  Identification rate for the combination of two and three sensors under different illumination conditions (NA = Natural, IR = Infrared, AR = Artificial)

| Sensors   | Normalization | Train | Test | | | |
|-----------|---------------|-------|------|---|---|---|
|           |               |       | NA   | IR | AR | |
|           |               |       | 3    | 4 | 3 | 4 | 3 | 4 |
| VIS&NIR   | NO            | NA 1&2| 91.7 | 96.1| 51.2| 58 | 91.7 | 92.7 |
| VIS&NIR   | YES           | NA 1&2| 94.1 | 96.6| 82  | 81.5| 94.6  | 94.6  |
| VIS&NIR   | NO            | IR 1&2| 93.7 | 90.7| 91.7| 92.7| 91.7  | 91.7  |
| VIS&NIR   | YES           | IR 1&2| 93.7 | 89.8| 97.6| 98  | 96.1  | 91.2  |
| VIS&NIR   | NO            | AR 1&2| 92.7 | 91.7| 49  | 52.2| 95.1  | 94.6  |
| VIS&NIR   | YES           | AR 1&2| 96.1 | 93.7| 86.8| 82.4| 97.07 | 96.1  |
| VIS&TH    | NO            | NA 1&2| 91.7 | 87.3| 85.9| 82.9| 95.1  | 88.8  |
| VIS&TH    | YES           | NA 1&2| 94.1 | 88.3| 91.7| 87.8| 95.1  | 88.3  |
| VIS&TH    | NO            | IR 1&2| 96.1 | 89.3| 90.2| 84.4| 96.6  | 90.2  |
| VIS&TH    | YES           | IR 1&2| 96.6 | 86.8| 97.1| 94.1| 98    | 90.7  |
| VIS&TH    | NO            | AR 1&2| 95.6 | 89.3| 83.4| 82.9| 98    | 93.7  |
| VIS&TH    | YES           | AR 1&2| 95.6 | 89.3| 93.2| 91.2| 98    | 93.7  |
| NIR&TH    | NO            | NA 1&2| 91.7 | 96.1| 85.9| 74.1| 95.1  | 96.6  |
| NIR&TH    | YES           | NA 1&2| 94.1 | 97.6| 91.7| 82  | 97.1  | 97.1  |
| NIR&TH    | NO            | IR 1&2| 96.6 | 89.3| 96.1| 96.6| 92.2  | 87.8  |
| NIR&TH    | YES           | IR 1&2| 94.6 | 84.4| 98.5| 99  | 90.2  | 85.4  |
| NIR&TH    | NO            | AR 1&2| 93.7 | 94.6| 77.1| 67.8| 97.1  | 93.7  |
| NIR&TH    | YES           | AR 1&2| 96.1 | 96.1| 84.4| 80.5| 98.5  | 93.7  |
| VIS&NIR&TH| NO            | NA 1&2| 94.6 | 97.6| 94.1| 91.7| 97.1  | 97.1  |
| VIS&NIR&TH| YES           | NA 1&2| 94.6 | 97.6| 94.1| 91.7| 97.6  | 97.1  |
| VIS&NIR&TH| NO            | IR 1&2| 98.5 | 95.1| 98  | 97.6| 98.5  | 97.1  |
| VIS&NIR&TH| YES           | IR 1&2| 98   | 94.6| 98.5| 100 | 99.5  | 96.6  |
| VIS&NIR&TH| NO            | AR 1&2| 97.1 | 95.1| 80.5| 74.6| 99.5  | 97.1  |
| VIS&NIR&TH| YES           | AR 1&2| 98.5 | 98.5| 93.7| 91.7| 99.5  | 98    |

**Fig. 16** Trained rule combining VIS and NIR classifiers for NA illumination for training and testing, session 4
Thus, near-infrared images are ignored and thermal images are weighted two times more than visible ones. This is reasonable considering the identification rates of each sensor alone (see Table 4: VIS = 60 %, NIR = 21 %, TH = 78 %). Using these optimal combination values, the identification rate reaches 84.9 %.

Figure 18 shows the contour plots as well as the maximum identification rate for the VIS, NIR and TH combination from top down and left to right for the following training and testing illumination conditions: NA–NA, NA–IR, NA–AR, IR–NA, IR–IR, IR–AR and AR–NA, AR–IR, AR–AR for session 4 and unnormalized feature vectors. Figure 19 represents the experimental results under the same illumination conditions for the normalized feature vectors case.

Conclusions

In this paper, a new face database has been presented. To the best of our knowledge, this is the first multisession database that consists of visible, near-infrared and thermal images acquired simultaneously and under different illumination conditions (natural, near-infrared and artificial).

The main conclusions about face recognition using a single sensor are as follows:

- The three studied sensors can provide good identification rates.
- The highest identification rate has been obtained for the NIR sensor under NIR illumination conditions.
- Thermal sensor is more stable along different illumination mismatches, as expected, and it also provides good enough identification rates and requires smaller number of coefficients. In addition, optimal feature selection is less critical than for the other sensors.
- On average, visible sensor provides higher identification rates.

The main conclusions when fusing two or three sensors are as follows:

- The combination improves the identification rates. The best system alone provides a 95.1 % identification rate, and the combined system reaches the 100 % in a particular scenario.
Fig. 19 Contour plots when combining VIS, NIR and TH sensors under the following training and testing illumination conditions: NA–NA, NA–IR, NA–AR, IR–NA, IR–IR, IR–AR and AR–NA, AR–IR, AR–AR for session 4 and normalized feature vectors

- In general, the three sensors are almost equally important, because a quite balanced weighting factor is obtained by exhaustive trial and error of the whole set of weighting combinations.
- Normalized feature vectors always outperform the unnormalized system for the trained combination rule and are slightly worse in 3 of 18 cases for the fixed combination rule.
- When studying the three sensors simultaneously, we have not found any couple of redundant sensors. The combined system takes advantage of the three spectral bands. In addition, the combined system is more robust with respect to illumination mismatch.

Acknowledgments This work has been supported by FEDER, MEC, TEC2009-14123-C04-04, KONTAKT-ME 10123, SIX (CZ.1.05/2.1.00/03.0072), CZ.1.07/2.3.00/20.0094 and VG20102014033.

References

1. Faundez-Zanuy M. Biometric security technology. IEEE Aerosp Electron Syst Mag. 2006;21(6):15–26.
2. Zhao W, Chellapa R, editors. Face processing: advanced modeling and methods, 1st ed. Academic Press; 2005. http://store.elsevier.com/Face-Processing-Advanced-Modeling-and-Methods/isbn-9780128845207/
3. Espinosa-Duró V, Faundez-Zanuy M, Mekyska J. Beyond cognitive signals. Cogn Comput. 2011;3:374–81. Springer.
4. Espinosa-Duró V, Faundez-Zanuy M, Mekyska J, Monte E. A criterion for analysis of different sensor combinations with an application to face biometrics. Cogn Comput. 2010;2(3):135–41.
5. Light L, Kayra-Stuart P, Hollander S. Recognition memory for typical and unusual faces. J Exp Psychol Hum Learn Mem. 1979;5:212–28.
6. Brennan SE. The caricature generator. Leonardo. 1985;18:170–8.
7. Hizem W, Allano L, Melfakh A, Dorizzi B. Face recognition from synchronised visible, near-infrared images. IET Signal Process. 2009;3(4):282–8.
8. Socolinsky DA, Selinger A. Thermal face recognition in an operational scenario. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition (CVPR’04). Vol. 2. 2004. pp. II-1012–II-1019.
9. Viola P, Jones M. Robust real-time object detection. Technical Report CRL 2001/01, Cambridge Research Laboratory, 2001.
10. Mekyska J, Espinosa-Duró V, Faundez-Zanuy M. Face segmentation: a comparison between visible and thermal images. In: IEEE 44th international Carnahan conference on security technology ICCST 2010, San José, USA, 5–8 Oct 2010.
11. Faundez-Zanuy M, Roure-Alcocé J, Espinosa-Duró V, Ortega JA. An efficient face verification method in a transformed domain. Pattern Recogn Lett. 2007;28(7):854–8. Elsevier.
12. Turk M, Pentland A. Eigenfaces for recognition. Journal Cognitive Neuroscience. 1991;3(1):71–86. Massachusetts Institute of Technology.
13. Jain AK. Fundamentals of digital image processing. New York: Prentice Hall; 1989.
14. Vivaracho C, Faundez-Zanuy M, Gaspar JM. An efficient low cost approach for on-line signature recognition based on length normalization and fractional distances. Pattern Recogn. 2009;42(1):183–93. Elsevier.
15. Mekyska J, Faundez-Zanuy M, Smekal Z, Fabregas J. Score fusion in text-dependent speaker recognition systems. Lect Notes Comput Sci. 2011;6800:120–32.
16. Franois D, Wert V. The concentration of fractional distances. IEEE Trans Knowl Data Eng. 2007;19(7):873–86.
17. Faundez-Zanuy M. Data fusion in biometrics. IEEE Aerosp Electron Syst Mag. 2005;20(1):34–8.
18. Kwon OK, Kong SG. Multiscale fusion of visual and thermal images for robust face recognition. In: IEEE international conference on computational intelligence for homeland security and personal safety. Apr 2005. pp. 112–116.
19. Moon S, Kong SG, Yoo JH, Chung K. Face recognition with multiscale data fusion of visible and thermal images. In: IEEE
20. Bhowmik MK, Bhattacharjee D, Nasipuri M, Basu DK, Kundu M. Classification of fused images using radial basis function neural network for human face recognition. In: IEEE 2009 world congress on nature & biologically inspired computing (NaBIC 2009). 2009. pp. 19–24.

21. Bhowmik MK, Bhattacharjee D, Nasipuri M, Basu DK, Kundu M. Optimum fusion of visual and thermal face images for recognition. In: 2010 IEEE sixth international conference on information assurance and security. 2010. pp. 311–316.

22. Singh R, Vatsa M, Noore A. Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. Pattern Recogn. 2008;41:880–93.

23. Neagoe VE, Ropot AD, Mugioiu AC. Real time face recognition using decision fusion of neural classifiers in the visible and thermal infrared spectrum. In: IEEE conference on advanced video and signal based surveillance. 2007. pp. 301–306.

24. Pop FM, Gordan M, Florea C, Vlaicu A. Fusion based approach for thermal and visible face recognition under pose and expressivity variation. In: 9th RoEduNet IEEE international conference. 2010. pp. 61–66.

25. Buyssens P, Revenu M. Fusion levels of visible and infrared modalities for face recognition. In: 2010 fourth IEEE international conference on biometrics: theory applications and systems (BTAS). 2010. pp. 1–6.

26. Raghavendra R, Dorizzi B, Rao A, Kumar GH. Particle swarm optimization based fusion of near infrared and visible images for improved face verification. Pattern Recogn. 2011;44:401–11.

27. Arandjelovic O, Hammoud R, Cipolla R. Thermal and reflectance based personal identification methodology under variable illumination. Pattern Recogn. 2010;43:1801–13.