Pattern Anal Applic
DOI 10.1007/s10044-015-0493-z

INDUSTRIAL AND COMMERCIAL APPLICATION

Periocular recognition: how much facial expressions affect performance?

Elisa Barroso[1] · Gil Santos[1] · Luis Cardoso[1] · Chandrashekhar Padole[1] · Hugo Proença[1]

Received: 19 March 2013 / Accepted: 7 June 2015

Abstract Using information near the human eye to perform biometric recognition has been gaining popularity. Previous works in this area, designated *periocular recognition*, show remarkably low error rates and particularly high robustness when data are acquired under less controlled conditions. In this field, one factor that remains to be studied is the effect of facial expressions on recognition performance, as expressions change the textural/shape information inside the periocular region. We have collected a multisession dataset whose single variation is the subjects' facial expressions and analyzed the corresponding variations in performance, using the state-of-the-art periocular recognition strategy. The effectiveness attained by different strategies to handle the effects of facial expressions was compared: (1) single-sample enrollment; (2) multisample enrollment, and (3) multisample enrollment with facial expression recognition, with results also validated in the well-known *Cohn–Kanade AU-Coded Expression* dataset. Finally, the role of each type of facial expression in the *biometrics menagerie* effect is discussed.

✉ Hugo Proença
hugomcp@di.ubi.pt

Elisa Barroso
ebarroso@di.ubi.pt

Gil Santos
gsantos@di.ubi.pt

Luis Cardoso
lcardoso@di.ubi.pt

Chandrashekhar Padole
cpadole@di.ubi.pt

[1] Department of Computer Science, IT-Instituto de Telecomunicações, University of Beira Interior, Covilhã 6200, Portugal

Keywords Periocular recognition · Biometrics

1 Introduction

Using the periocular region to perform biometric recognition has recently gained popularity. By acquiring a region that is similar to that used by iris recognition systems, the key insight is to use not only the discriminating information inside the iris, but also all of the textures from the skin near the eye as well as the shape of the eyelid, the eyebrow and the eyelashes. In this area, various methods have recently been proposed, including the most relevant from Park et al. [1], which characterized the periocular texture using local binary patterns (LBP), histograms of oriented gradients (HOG) and scale-invariant feature transform (SIFT). A subsequent work [2] described additional factors that affect performance, including segmentation inaccuracies, partial occlusions and pose. Lyle et al. [3] classified gender and ethnicity based on periocular data, using LBP features to feed a support vector machine. A noteworthy conclusion was that the effectiveness is comparable to that obtained using the entire face. Woodard et al. [4] studied the effect of fusion techniques on periocular and iris data in non-ideal scenarios, concluding that fusion at the score level improves performance. Bharadwaj et al. [5] used visible light data and fused a global matcher (spatial envelope) to circular linear binary patterns. More recently, Ross et al. [6] handled non-ideal ocular data and discussed the challenges around sample deformation and varying illumination, using probabilistic deformation models and maximum-a-posteriori estimation filters to fuse descriptors. Hollingsworth et al. [7] compared the recognition ability of humans and machines using periocular data, concluding that automated strategies have at least as much

Published online: 25 June 2015

effectiveness as humans. Focusing on robustness, Woodard et al. [8] represented the skin texture and color using separate features, fusing both types of information. Finally, Crihalmeanu and Ross [9] fused periocular recognition techniques to methods that describe the sclera textures and vasculature patterns.

In this paper, we are particularly interested in the effect of facial expressions on the effectiveness of periocular recognition. As Fig. 1 illustrates, in no other part of the human body has as many muscles interact as in the face, which gives insight into our experiments. We objectively assess performance variations that result from varying facial expressions compared with considering exclusively neutral data. We have collected a multisession dataset, in which the main variation factor is precisely the subjects' facial expressions. We analyze the role of facial expressions in the well-known *biometric menagerie* effect, aiming to perceive how much they contribute to classifying a given subject in a menagerie family. The main contributions of this paper are as follows:

- Announcing a new dataset (*faceExpressUBI*) that contains multisession data from 184 subjects; the main varying factor is the subjects' facial expressions. As detailed in the annex, this dataset is freely available to the research community and constitutes a valuable resource for analyzing biometric effectiveness in either facial or periocular recognition when dealing with facial expressions;

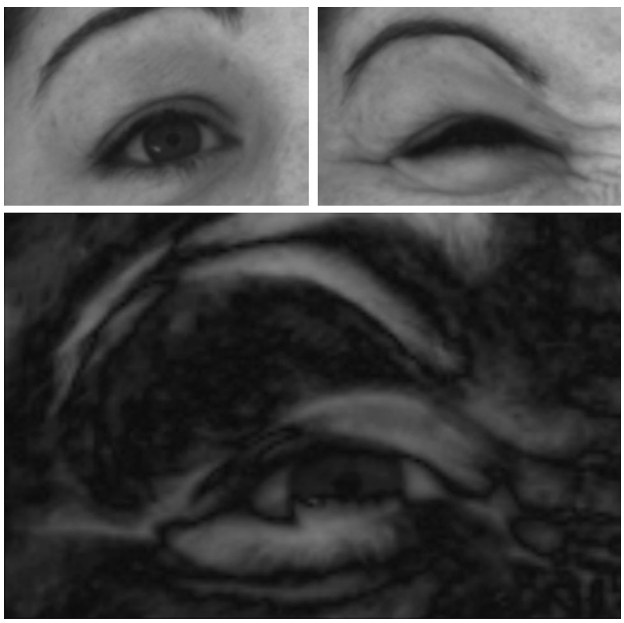


Fig. 1 Differences in texture and shape inside the periocular region due to varying facial expressions. The *top-left* image regards a neutral expression and the *top-right* image has features a *happy* expression. Image at the *bottom* shows the differences between the aligned images, where brightest corresponds to most notorious changes

- Quantifying the performance decreases that likely occur in periocular recognition due to the effect of facial expressions on both gallery and probe data.
- Assessing the effectiveness of three strategies to handle facial expressions: (1) single-sample neutral gallery data, (2) multisample gallery data, and (3) recognizing facial expressions in the biometric process.
- Identifying facial expressions that more likely contribute to including a given subject in each family defined in the *biometric menagerie* effect.

All research work reported in this paper can be reproduced, with the required information given in the annex. We describe how to access the data, the source code and the sets of pairwise comparisons performed for each experiment.

The remainder of this paper is organized as follows. Section 2 summarizes the datasets related to the work in this paper. Section 3 gives a detailed description of the *FaceExpressUBI* dataset. Section 4 reports our experiments and discusses the results. Finally, Sect. 5 presents the conclusions.

2 Facial expressions datasets

2.1 Related datasets

The literature describes many facial datasets. In Table 1, we summarize the most important features of each: the number of subjects, facial expression types and the existence of annotation data.

The Cohn–Kanade database [10] contains 504 image sequences of facial expressions from 100 subjects, ranging in age from 18 to 30 years. Apart from neutral faces, six types of expressions were identified: joy, surprise, anger, fear, disgust, and sadness. In a subsequent version (the extended Cohn–Kanade database), the number of sequences is increased by 22 number of subjects, by 27 database [12] contains over 1100 sequences with over 150 actions from 24 mostly Caucasian subjects. The Radboud dataset [13] (*RaFD*) includes 67 Caucasian subjects displaying eight emotions: anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. Each emotion is shown with three different gaze directions from five camera angles [14]. The FACES dataset comprises 171 Caucasian subjects of varying ages, displaying six different facial expressions [15]. The FEED dataset has 18 different subjects displaying six basic emotions and includes head movements in different directions. The expressions include happiness, disgust, anger, fear, sadness, surprise, and neutral [13]. The JAFFE database has ten people and six basic expressions, for 219 total still images, acquired in a heavily controlled environment [16]. The MMI dataset contains 52 subjects, ranging from 19 to 62 years and is one of the few

Table 1 Summary of the most relevant facial datasets. {N, U, A, C, D, Z, F, H, M, S, R} denote, respectively, neutral, smile, anger ,scream, disgust, sleepy, fear, happy, open mouth, sad and surprise expressions

Name	Years	Subj.	Imgs.	Resol.	Expressions	Obs.
Ekman-Hager	1996	24	1100	360 × 240	H, S, R, A, D, F	Luminance was normalized
FERET	1996	1199	14,051	256 × 384	N, U	–
University of Maryland	1997	40	2800	560 × 240	H, S, R, A, D, F	–
Yale face	1997	15	165	320 × 243	H, N, S, R, Z and wink	–
JAFFE	1998	10	219	256 × 256	N, H, S, R, A, D, F	Without occlusions, only Japanese female models
AR	1998	126	40,000	768 × 576	N, U, A, C	Illumination: left, right and both light on; occlusions: sunglasses and scarf
CK	2000	100	2300	640 × 490	H, S, R, A, D, F	Frontal and 30° to the right direction; Illumination: reflective umbrellas, ambient lighting, single and dual high intensity lamps
CK+	2000	123	2829	640 × 490	H, S, R, A, D, F	Frontal and 30° to the right direction; Illumination: reflective umbrellas, ambient lighting, single and dual high intensity lamps
CMU PIE	2000	68	41,368	640 × 480	N, U, Z	13 synchronized high-quality color cameras and 21 flashes
Notre dame humanID	2002	>300	>15,000	1600 × 1200	N, U	Lighting configurations: right, left and frontal illumination
CAS-PEAL	2003	377	30–900	360 × 480	N, U, A, R, Z, open mouth	9 cameras around the subject in a semicircular distribution
KFDB	2003	1000	52,000	640 × 480	N, H, R, A, Z	8 lights located around the subject at 45° and 15° intervals
RU-FACS-1	2004	100	400–800 min	720 × 576	False and truth opinion	Synchronized digital video from 4 video cameras
Equinox infrared	2005	91	14,560	240 × 320	U, A, R	8–12 μm spectral range and visible
MMI	2005	52	4108	720 × 576	H, S, R, A, D, F	–
University of Texas	2005	284	10 min with 11 emotions per person	720 × 480	N, H, S, R, A, D, F, U, boredom, disbelief, puzzlement	–
BU-3D FE	2006	1000	2500	1040 × 1329	N, H, S, R, A, D, F	Images and 3D models
FEED	2006	18	399	320 × 240	N, H, S, R, A, D, F	–
Spontaneous expressions	2007	28	112	–	N, H, R, D	–
BU-4D FE	2008	101	60,600	1040 × 1320	H, S, R, A, D, F	–
Radboud faces	2010	67	8040	1024 × 681	N, H, S, R, A, D, F, contempt	3 gaze directions, 5 different camera angles, three 500-W flashes were used
SAVEE	2011	4	–	–	H, S, R, A, D, F, N	Frontal faces were painted with 60 markers

multiracial datasets. Images are classified into two classes: posed and spontaneous expressions [17]. The Rochester/UCSD Facial Action Coding System Database (*RU—FACS*—1) contains 100 subjects, with approximately 2.5 min of video recorded for each. The opposition paradigm is used to acquire the images, wherein the subjects are queried about some issue and asked to take the opposite stand from what they previously reported. The singularity of the Surrey Audio-Visual Expressed Emotion Database (*SAVEE*) is that each subject's face is painted with 60 markers, although only for four English males [18, 19]. The Spontaneous Expressions Database contains 28 subjects and 112 total images, and four expressions are considered: joy, surprise, disgust, and neutral [20]. The AR dataset

contains images of 126 individuals, 70 men and 56 women, recorded in a frontal pose twice at a 2-week interval. Four facial expressions are considered: neutral, smiling, angry, and screaming. It is characterized by occlusions, including sunglasses, sunglasses/left light, sunglasses/right light, scarf, scarf/left light, and scarf/right light [18]. (*CAS-PEAL*) is a large-scale Chinese face database with variations in illumination, position and expressions. It contains 99,594 images of 1040 individuals, classified into 5 expressions. Similarly, the Carnegie-Mellon PIE database contains 3 facial expressions acquired from 13 synchronized color cameras, for 41,368 total images from 68 individuals [21]. The Equinox Infrared Face Database was collected by two synchronized sensors using long-wave infrared radiation and visible light imagery. The resulting image pairs are co-registered to within 1/3 of a pixel. For each subject, a 4-s (40 *frames*) video sequence is recorded while the subject pronounced the vowels. It also considers 3 facial expressions: smiling, frowning, and surprise [22]. As part of the FERET program, a dataset was collected in 15 sessions from 1199 individuals. During each acquisition session, 13 conditions with varying facial expressions, illumination and occlusion were captured [23]. The resulting changes in facial expression are typically subtle, often switching between neutral and smiling. The Korean face dataset contains 1000 facial subjects, collected in the middle of an octagonal frame carrying seven cameras and eight light sources against a blue screen background, considering 5 facial expressions: neutral, happy, surprise, anger, and blinking [24]. At Notre Dame University, a dataset from over 300 subjects was collected, and two facial expressions were considered: neutral and smiling [22]. At the University of Texas, a large database of static and video clips of faces was collected from 284 subjects, mostly Caucasians between 18 and 25 years old. Subjects were imaged at close range, and happiness, sadness, fear, disgust, anger, puzzlement, laughter, surprise, boredom, or disbelief expressions were considered [25]. At the University of Maryland, a dataset of 40 subjects of diverse racial and cultural backgrounds was collected at full frame rate while asking subjects to display their own choice of expressions. The occurrences of the six basic emotions were not balanced, with happiness, surprise, disgust, and anger appearing more frequently than sadness and fear [22]. The Yale Face Database contains 165 images of 15 subjects, in a variety of conditions, including with and without glasses, illumination variation, and changes in facial expressions, e.g., happy, normal, sad, sleepy, surprised, and winking [22]. The Binghamton University 3D Facial Expression Database includes 100 subjects, each one performing 7 expressions Ñ neutral, happiness, disgust, fear, anger, surprise, and sadness Ñ and four levels of intensity [26]. Finally, the BU-4D FE Database comprises 58 female and 43 male subjects, with a variety of ethnic/racial ancestries, including Asian, Black, Hispanic/Latino, and White, for 606 3D facial expression sequences captured from 101 subjects. Further, each model of a 3D video sequence has a resolution of approximately 35,000 vertices. It includes the salient features of the 3D database in the previous subsection and its dynamic characteristics [13].

3 FaceExpressUBI dataset

As described above, though there are many datasets reported in the literature, several drawbacks have been detected that led us to acquire a new one: 1) the FaceExpressUBI dataset has higher resolution (2056 x 2452 pixels) than most freely available datasets; 2) it has a single variation factor, differing from most of those described above; and 3) each image is associated with an annotation file using the following meta-data: coordinates for the face, mouth, nose, periocular region, eye centers and eyeglasses. We have also manually selected a data subset, designated *keyframes*, in which the facial expressions considered (e.g., neutral, angry, fear, disgust, happy, sad and surprised) are most evident.

3.1 Imaging framework and setup

The imaging framework was installed in five different places under both natural and artificial lighting sources. We used several marks on the floor to mark the video camera location and subjects' position, at a distance of 1.2 m from the camera. Each participant was asked to perform seven expressions, and each expression was recorded for approximately 5 s at a frame rate of 7 fps. At a minimum, two acquisition sessions were performed per subject, with a minimum interval of two weeks between sessions. The dataset contains 184 participants: volunteers were 10–48 years of age, 35 % female, 93 % Caucasian, 3 % Latino, 1 % Asian and 3 % African, and 12 % of the participants wore eyeglasses. We used an AVT Stingray F-504B camera with a resolution of 2452 $\times$ 2056, and collected 90,160 total images (Table 2). Figure 2 illustrates a sequence of images that comprises the neutral expression plus six facial expressions, with the type of expression and the frame index denoted below each image.

To address the issue of whether experiments performed in this dataset produce statistically significant results, we consider ρ as the classifier error rate, α as the confidence interval and $\hat{P}$ as the error rate estimated over a finite number of test patterns. At an α-confidence level, the true error rate should not exceed the estimated error rate by an

Table 2 Summary of the FaceExpressUBI dataset images, of the image acquisition framework and setup and of the subjects who offered themselves as volunteers to the imaging sessions

Image acquisition framework and setup	Characterization
Camera	AVT stingray F-504B
Color representation	Black and white
Shutter speed	47 (min)/67,000,000, auto shutter
Total pixels	5 Megapixels
Frame rate	7 fps
Focal length	35 mm
Cell size	3.45 μm x 3.45 μm
Focal length	35 mm

Resultant images	Details
Format	Tiff
Bit depth	8 bit
Vertical resolution	2056 pixels
Horizontal resolution	2452

Volunteers	Characterization
Totals	184 subjects; 90,160 images; 490 images per subject
Gender	65 females-35 %; 119 males-65 %
Ethnicity	93 % Caucasian Europeans; 3 % American Latin; 3 % Africans; 1 % Asians
Age	[0; 20]–33.9 %; [21; 25]–48 %; [26; 30]–9.3 %; [31; 35]–4.4 %; [36; 99]–4.4 %

amount larger than $\varepsilon(N,\alpha)$.Guyon et al. [27] fixed ε to be a given fraction of P ($\varepsilon(N,\alpha)=\beta P$).They considered that recognition errors are Bernoulli trials, concluding that the number of trials N required to achieve a $(1-\alpha)$ confidence in the error rate estimated is given thus:

$$N=-\ln(\alpha)/\left(\beta^2 P\right), \tag{1}$$

having authors obtained typical values for α and $\beta(\alpha=0.05;\beta=0.2)$ and recommended a simpler equation:

$$N\approx(100/P). \tag{2}$$

Assuming that each frame is used to generate a biometric template, the remaining frames from the same subject are used to analyze genuine variability, and the remaining ones from different faces are used to analyze impostor variability. We obtain a bound for the error to test using statistical significance. The 90,160 images of the FaceExpressUBI dataset enable 22,089,200 genuine and 4,042,323,600 impostor comparisons. This guarantees statistical significance in experiments with an empirical error rate $\hat{P}$ as low as 2.474×10^{-8} percent, which should clearly be considered a lower bound.

3.2 Annotation data

For each image in the dataset, several regions of interest were automatically detected based on Haar classifier cascades. A human observer also validated the results. For each image, the resulting annotation file contains six lines (Fig. 3); the first four lines give the coordinates (upper-left and bottom-right corners) for the face, periocular region, nose and mouth. The fifth line gives the center coordinates for the right and left eyes. The last line expresses *key-frames*, i.e., the frames in which the corresponding facial expression is most evident. Eyeglasses are also marked.

4 Experiments

The results given here consider the recognition method proposed by Park et al. [2]: the authors used a Haar cascade to detect faces and heuristic rules based on human face anthropometry to define the periocular region of interest. This region was described using three texture encoding strategies: histograms of oriented gradients (HOG), local binary patterns (LBP) and scale-invariant feature transform (SIFT). Images were aligned and normalized for both scale and translation according to the annotation data, from which LBP, HOG and SIFT descriptors were extracted. In matching HOGs and LBPs, the χ^2 distance was used, whereas for SIFT, the distance-ratio criterion suggested by Lowe was applied. Fusing the distance values between descriptors was performed using logistic regression [28], which is equivalent to a single-output neural network with an activation function trained under log loss:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \quad (3)$$

where the fraction $p/(1-p)$ is called *the odds* of a positive match, i.e., the ratio between that probability and its complement. β_i values are weights relating the distances between descriptors x_i (HOG, LBP and SIFT) to the odds.

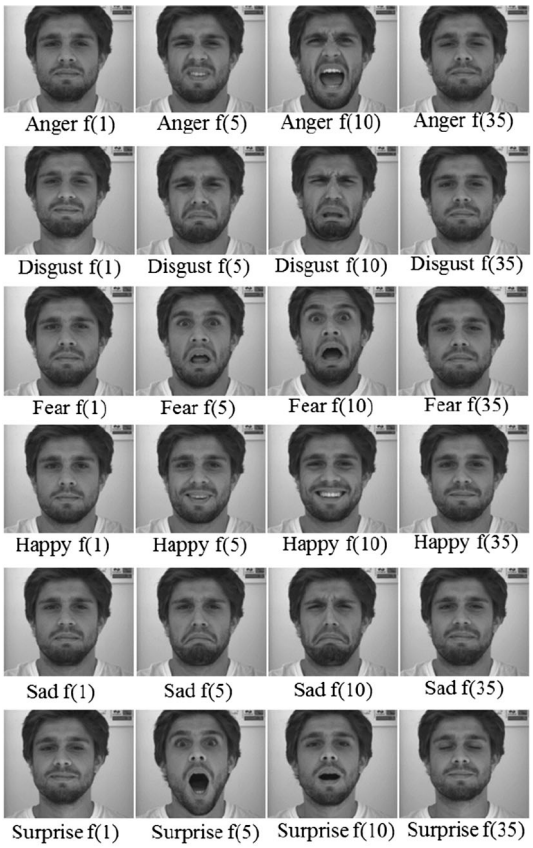


Fig. 2 Examples of the images of *FaceExpressUBI* dataset, considering the 1st, 5th, 10th and 35th frames per expression

Our experiments were divided into four main sections: (1) quantifying the performance decreases due to facial expressions; (2) studying the linear correlations between expressions; (3) analyzing performance using different enrollment/recognition strategies to handle facial expressions; and (4) assessing the role of facial expressions in the biometrics menagerie effect, according to the proposal of Yager and Dunstone [29].

4.1 Datasets

All the experiments reported in this section were carried out using two datasets: (1) the above described *faceExpressUBI*, which was used as main source; and (2) the *Cohn–Kanade AU-Coded Expression* [11] set, which served for validation purposes, i.e., to confirm that the results obtained in our dataset were statistically relevant. In Fig. 4 we give some examples of the both sets considered, where the images in the upper row regard the *faceExpressUBI* set, and the bottom row contains examples of the *Cohn–Kanade AU-Coded Expression* set, for which we considered exclusively frontal images. When comparing both sets the main difference regards the average data resolution, which is far higher in ours dataset. Both sets contain grayscale images, taken in relatively uncontrolled lighting conditions, with subjects standing frontal to the camera.

4.2 Performance analysis

To objectively measure the performance decreases due to facial expressions, the *faceUBIExpress* dataset was divided into twenty-nine subsets (each containing 17,934 genuine and 17,934 impostor comparisons), in which comparisons appear in the exact same order; only the facial expressions vary. Let the seven types of facial expressions (e.g., neutral, anger, disgust, fear, happiness, sadness and surprise) be denoted by $\mathbb{F}_i, i \in \{1, \ldots, 7\}$ and $(\mathbb{F}_i \leftrightarrow \mathbb{F}_j)$ denote the

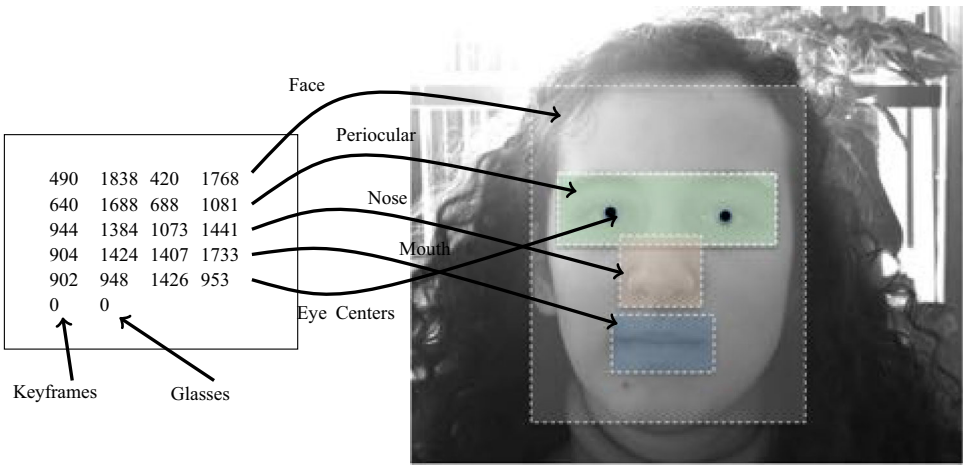
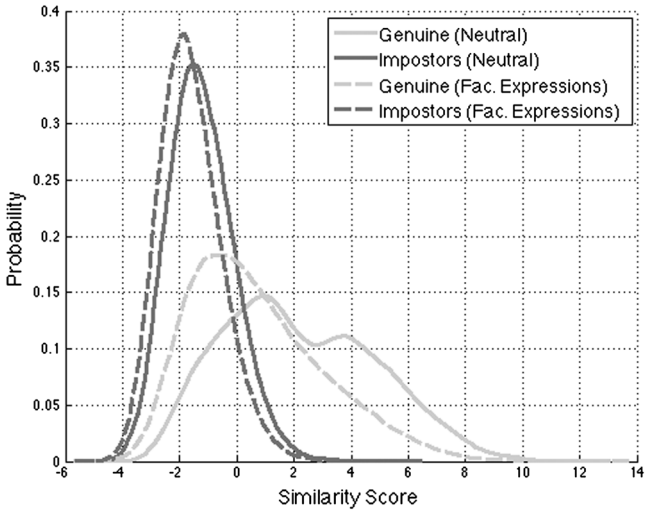


Fig. 3 Example of an annotation file, with the corresponding labels for comprehensibility purposes. Image is *disgust_311_01_f_3.tiff*. Each *line* contains the initial and final coordinates (*columns* and *rows*) of a facial component. The *fifth* row contains the coordinates of both eye centers. The *bottom* row denotes keyframes (1 = yes) and eyeglasses (1 = yes)

Fig. 4 Examples of the data considered in the experiments reported in this section: the *upper row* contains samples of the *faceExpressUBI* set, whereas the *bottom row* regards the *Cohn–Kanade AU-Coded Expression* set



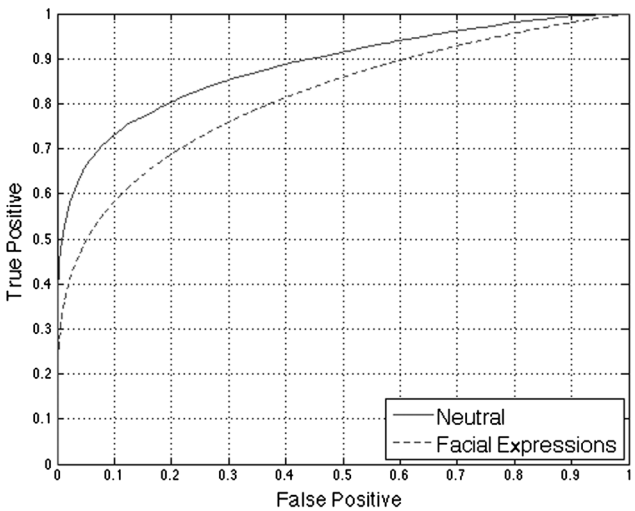


Fig. 5 Variations in recognition performance when considering only data of neutral facial expressions (*Neutral* labels) and data of varying facial expressions (*Fac. Expressions* labels). The AUC values decreased from 0.884 to 0.816 and the decidability d' from 1.679 to 1.263. Values regard the *faceUBIExpress* set

set of comparisons between elements with facial expression $\mathbb{F}_i$ (gallery) and $\mathbb{F}_j$ (probe). Figure 5 compares the genuine/impostor histograms (left plot) and ROC curves (right plot) obtained when considering exclusively neutral expressions ($\mathbb{F}_1 \leftrightarrow \mathbb{F}_1$) with varying expressions ($\mathbb{F}_i \leftrightarrow \mathbb{F}_j$, $\forall i,j \in \{1,\ldots,7\}$). The results are evident (AUC decreased from 0.884 to 0.816 and the decidability from 1.679 to 1.263). The histogram analysis shows that decreases are mostly due to genuine scores moving toward the impostor distribution, a typical occurrence when biometric systems handle data with degraded quality.

To perceive the individual effect of each expression type, Fig. 6 gives boxplots of the variations between the genuine matching scores of $\mathbb{F}_1 \leftrightarrow \mathbb{F}_1$ and ($\mathbb{F}_1 \leftrightarrow \mathbb{F}_i$, $\forall i = \{2,\ldots,7\}$). The median decrease is denoted by the horizontal solid lines in the center of each box, and the first and third quartile values are denoted by the top and bottom of the box marks. The upper and lower whiskers are denoted by the horizontal lines outside each box, and the outliers appear as cross points. All facial expressions decrease the matching scores of genuine comparisons, which agrees with the movement described in the left plot of Fig. 5. This is most obvious for the expression *disgust* and less so for *happiness*. The results are statistically validated using paired Student's t tests, as differences between corresponding matching scores are observed to be approximately normal. $t = \frac{\mu_1 - \mu_2}{S_{1,2}\sqrt{2/n}}$, where μ is the mean of scores obtained from neutral expressions and facial expressions data, S is the sample standard deviation and $n = 17,934$ is the dimension of samples. All tests consider the null hypothesis (H_0: *The facial expression does not decrease the matching scores, compared with data from the neutral expression*), which is clearly rejected at the $\alpha = 0.01$ significance level, with residual p values and 99 % confidence intervals for the difference of means of [1.107, 1.108] (anger), [1.260, 1.260] (disgust), [0.782, 0.783] (fear), [0.474, 0.474] (happiness), [0.735, 0.736] (sadness) and [1.014, 1.015] (surprise).

Figure 7 illustrates two extreme cases. On the left, the decrease in the matching score is maximal (−13.65), with

evident differences in the eyebrows, eyelids, and skin texture regions. Conversely, decreases are far lower for less expressive subjects, and they do not even occur in some cases, as illustrated by the case shown on the right, in which the matching score is even higher than the pairwise matching of neutral data of the same subject (7.63).

4.3 Correlation between facial expressions

Analyzing the linear correlation between scores obtained when matching data with varying facial expressions might be important at two levels: the probability of mismatching genuine comparisons due to expressions and the probability of false matches occurring in impostor comparisons due to facial expressions. Table 3 gives such results in (X, Y) format, where X is the correlation of genuine comparisons and Y that of impostor correlations. Values regard the *faceUBIExpress*, with bold values denoting the maximum/minimum levels of linear correlation observed between facial expressions. The most notable correlations are highlighted in bold font. The most correlated are neutral $\leftrightarrow$ happy for both genuine and impostor comparisons. The scores generated for anger $\leftrightarrow$ surprise are least correlated in the impostor comparison, which might have biological roots in the disjointed set of muscles evolved in the corresponding facial expressions. On average, significant correlation levels are observed: a mean of 0.62 and standard deviation of 0.02 for the genuine comparisons and a mean of 0.52 and standard deviation of 0.02 for the impostor.

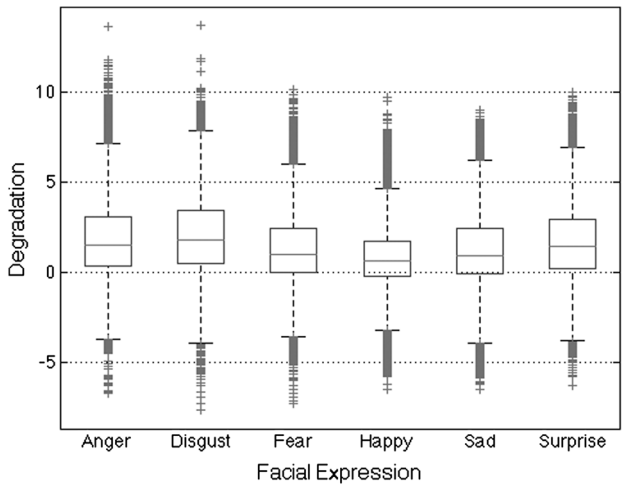


Fig. 6 *Boxplots* of the decreases in matchings scores when using probe data with different facial expressions against neutral gallery data, in comparison to scores obtained when both gallery and probe are neutral (degradation = 0). Results regard the *faceUBIExpress* set

4.4 Enrollment/recognition strategies

Having observed that facial expressions actually decrease the effectiveness of periocular biometrics, in this section we discuss different enrollment and recognition strategies to compensate for the effects of facial expressions. For validation purposes, the results are given here both for the *faceUBIExpress* and *Cohn–Kanade AU-Coded Expression* datasets.

4.4.1 Uncontrolled setup

In a totally uncontrolled setup, either gallery and probe data might have varying expressions if the effect of facial expressions is neglected. Let $\mathfrak{G} = \{g, \ldots, g_n\}$ and $\mathfrak{P} = \{p_1, \ldots, p_n\}$ be the gallery and probe sets, each containing n images, $g_i, p_i \in \{\mathbb{F}_1, \ldots, \mathbb{F}_7\}$. Figure 8 gives the performance in such conditions, using the results obtained when all expressions are neutral as the comparison term. In this case, the decreases are substantial and consistent across all performance ranges (AUC decreased from 0.884 to 0.815 and decidability from 1.679 to 1.328 in the *faceUBIExpress*, whereas a slightly larger decrease in performance was observed for the *Cohn–Kanade AU-Coded Expression* dataset), confirming that-indeed-periocular recognition effectiveness decreases for data with different facial expressions. A such, systems must handle the effect of facial expressions to optimize performance.

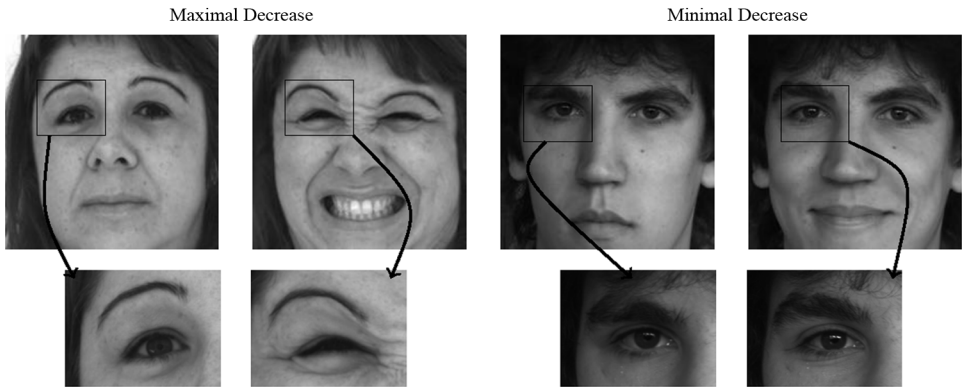


Fig. 7 Examples of two genuine images pairings of the *faceUBIExpress* set, where variations in the matching scores due to the effect of facial expressions attained extremes. The case shown at left regards the maximal decrease (−13.65), and at right we show a case where the matching score even improved (7.63), when compared to pairwise matching of data with neutral expression

Table 3 Levels of linear correlation between the matching scores observed when using data of varying facial expressions

	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise
Neutral	1.000, 1000	0.617, 0.536	**0.547**, 0.494	0.675, 0.577	**0.777**, **0.673**	0.676, 0.601	0.662, 0.533
Anger	–	1.000, 1000	0.639, 0.580	0.623, 0.501	0.621, 0.519	0.609, 0.529	**0.547**, **0.397**
Disgust	–	–	1.000, 1000	0.576, 0.484	0.578, 0.515	0.564, 0.515	0.582, 0.417
Fear	–	–	–	1.000, 1000	0.653, 0.538	0.668, 0.565	0.654, 0.536
Happy	–	–	–	–	1.000, 1000	0.670, 0.585	0.646, 0.514
Sad	–	–	–	–	–	1.000, 1000	0.631, 0.487
Surprise	–	–	–	–	–	–	1.000, 1.000

Results regard the *faceUBIExpress*

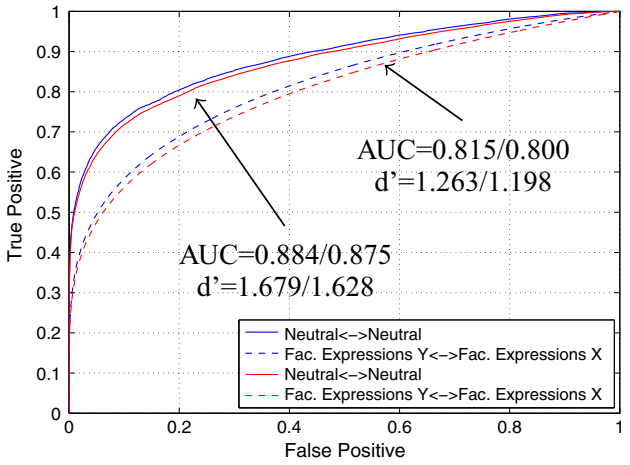


Fig. 8 Recognition performance in totally uncontrolled setups (Expression X ↔ Y), where either the probe and gallery data contains varying facial expressions, in comparison to considering only data of neutral expressions (Neutral ↔ Neutral). The *blue lines* regard the *faceUBIExpress* dataset, whereas the *red lines* correspond to results observed for the *Cohn–Kanade AU-Coded Expression* set. The AUC and decidability values are given in *faceUBIExpress*/*Cohn–Kanade* format

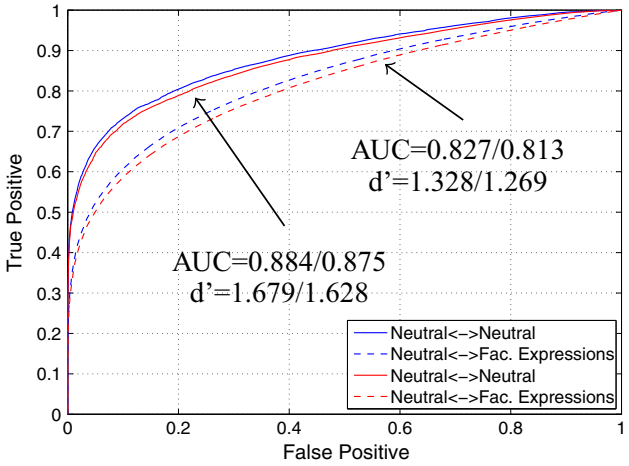


Fig. 9 Recognition performance in a semi-controlled setup, where gallery data is neutral and probes have varying facial expressions (Neutral ↔ Fac. Expressions), in comparison to considering only data of neutral expressions (Neutral ↔ Neutral). The *blue lines* regard the *faceUBIExpress* dataset, whereas the *red lines* correspond to results observed for the *Cohn–Kanade AU-Coded Expression* set. The AUC and decidability values are given in *faceUBIExpress* / *Cohn–Kanade* format

4.4.2 Single sample enrollment

In a slightly more controlled environment, we assess performance when matching neutral gallery data to probes of varying facial expressions, i.e., $g_i \in \mathbb{F}_1$ and $p_i \in \{F_1, \ldots, F_7\}$. Figure 9 gives the obtained results using the same reference values as above. In this case, though facial expressions can be considered as poorly handled (AUC=0.884 in neutral data and AUC = 0.827 with facial expressions for *faceUBIExpress*, and AUC = 0.875 in neutral data and AUC=0.800 in data with facial expressions for *Cohn–Kanade AU-Coded Expression* set), we observe a slight increase in the recognition effectiveness compared with the uncontrolled setup (AUC = 0.815). We thus concluded that gallery data with constant expressions might be the best choice if using single-sample enrollment. Based on the correlation values given in Table 3, *neutral* or *sad* expressions might be the best choices.

4.4.3 Multisample enrollment: one probe against all gallery

Using multisample gallery data, a possibility might be to enroll one sample per different expression. Each probe is then matched against all gallery expressions, and the minimal dissimilarity is used as a matching score for each identity, i.e., each probe p_i is matched against g_j, $j \in \{\mathbb{F}_1, \ldots, \mathbb{F}_7\}$, and the final score is given by $\min\{p_i \leftrightarrow g_j\}$. Figure 10 shows that the results obtained according to this strategy do not improve the recognition effectiveness at all; they lead to a decrease from the uncontrolled setup. Observing the genuine/impostor histograms indicates that this decrease is an evident movement of the impostor scores toward the genuine distribution, due to the *min* operator, which was observed both for the *faceUBIEXpress* and the *Cohn–Kanade AU-Coded Expression* sets. We thus conclude that this strategy is not suitable for handling the

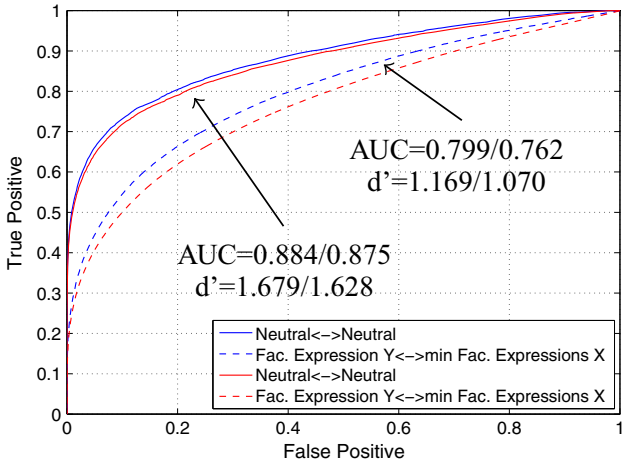


Fig. 10 Recognition performance when using multisample enrollment data, each one with a different facial expression (Expression X ↔ min Expression Y), in comparison to considering only data of neutral expressions (Neutral ↔ Neutral). The *blue lines* regard the *faceUBIExpress* dataset, whereas the *red lines* correspond to results observed for the *Cohn–Kanade AU-Coded Expression* set. The AUC and decidability values are given in *faceUBIExpress* / *Cohn–Kanade* format

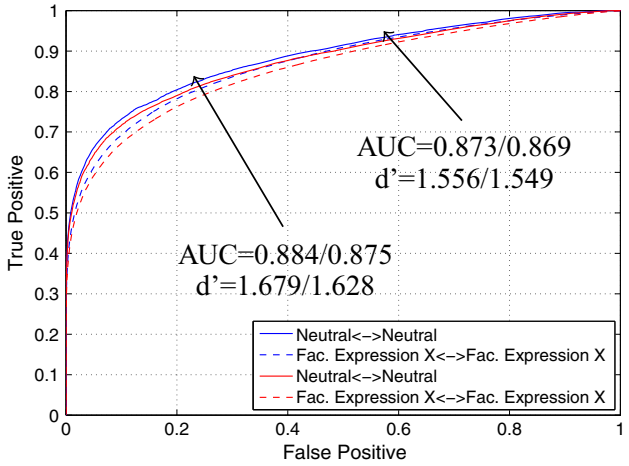


Fig. 11 Recognition performance obtained when recognizing the facial expression of each probe and matching only to gallery data of the same expression (Expression X ↔ Expression X), in comparison to considering only data of neutral expressions (Neutral ↔ Neutral). The *blue lines* regard the *faceUBIExpress* dataset, whereas the *red lines* correspond to results observed for the *Cohn–Kanade AU-Coded Expression* set. The AUC and decidability values are given in *faceUBIExpress* / *Cohn–Kanade* format

effect of facial expressions and might even be worse than simply neglecting expressions.

4.4.4 *Multisample enrollment: recognizing facial expressions*

Finally, the performance obtained when using a module for recognizing facial expressions is assessed, which corresponds to having gallery data from different expressions, recognizing the facial expression for each probe and only matching it against gallery data with the corresponding expression. This is equivalent to $p_i \leftrightarrow g_i$ and $(p_i, g_i \in \mathbb{F}_j, j \in \{1, \ldots, 7\})$. Figure 11 gives the results, maintaining the performance attained using exclusively neutral data as a comparison term. In this case, the decrease is minimal (AUC 0.884 to 0.873 and decidability d' from 1.679 to 1.556 in the *faceUBIExpress*, with a highly similar decrease in the *Cohn–Kanade AU-Coded Expression* set). Hence, we conclude that recognizing facial expressions before the biometrics process handles the effect of facial expressions almost perfectly if multisample gallery data of varying expressions are available.

4.5 Biometric menagerie

As Yager and Dunstone [29] suggested, not all subjects perform similarly in terms of biometric system recognition effectiveness. Various groups are labeled by animal names that reflect their discriminating features: *goats* are especially difficult to match, whereas *lambs* and *wolves* are characterized by their easiness in being matched with others. *Chameleons* have high matching scores for both genuine and impostor comparisons, and *phantoms* are the opposite: they tend to produce low matching scores against both themselves and others. *Doves* are the best possible case of biometric systems and match well against themselves and poorly against others. Finally, the worst type of conceivable subjects are *worms*, which are difficult to match against themselves but easy to match against others. We use a variant of the biometric menagerie index (BMI) suggested by Poh and Kittler [30] to characterize the extent of the menagerie effect, and we estimate the bias term instead of its square and the standard deviation instead of the total variance, obtaining an index in the [−1, 1] interval that, apart from magnitude, gives information about the deviation direction using global means (the original BMI index is in the [0,1] interval and concerns only the magnitude of the menagerie effect). Let $y_{ij}^k \in \mathbb{R}$ be the matching scores on class k (genuine or impostor) and j be the factor to be analyzed ($j \in \{1, \ldots 184\}$ subjects, or $j \in \{1, \ldots, 7\}$ facial expressions, or $j \in \{1, \ldots, 7 \times 184\}$ when analyzing the joint effect). The global and factor-specific j means are given thus:

$$\hat{\mu^k} = \frac{1}{JN^k} \sum_{j=1}^{J} \sum_{i=1}^{N^k} y_{ij}^k$$

$$\hat{\mu_j^k} = \frac{1}{N^k} \sum_{i=1}^{N^k} y_{ij}^k,$$

where $J = 7 \vee J = 184 \vee J = 1288$ and there are N^k scores

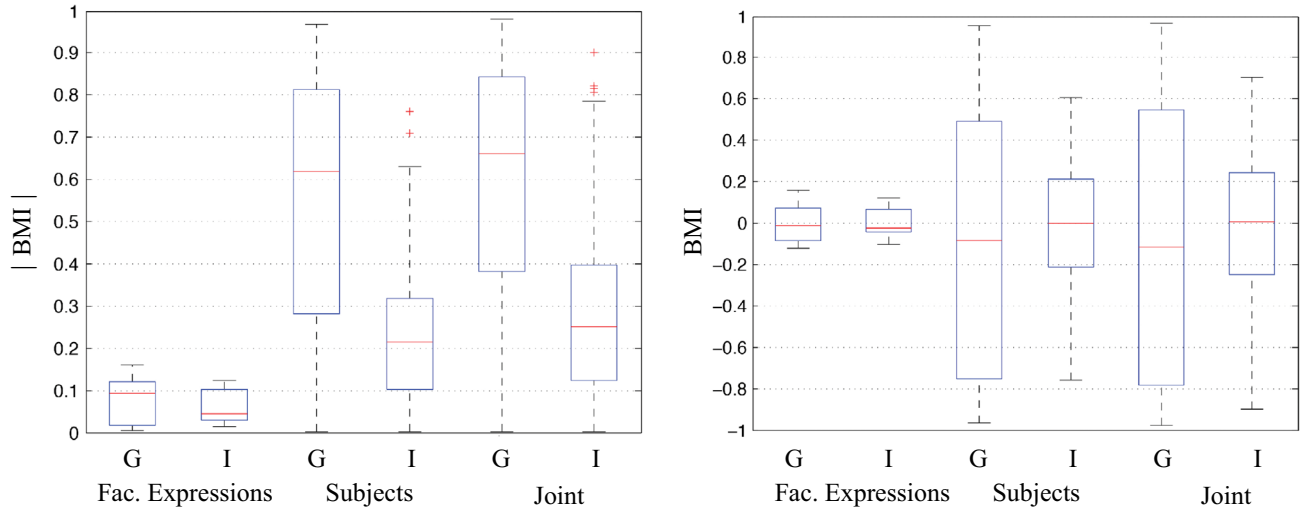


Fig. 12 *Boxplot* of the biometric menagerie indexes (BMI) obtained, with respect to facial expressions, subjects and joint effect, observed for the *faceUBIEXpress* and *Cohn–Kanade AU-Coded Expression* sets. The role of subjects on the menagerie effect appears to be stronger than the effect of facial expressions. Even though, facial expressions appear to highlight the menagerie potential for most subjects

per class k. The bias $\hat{V}_{kj}^{B}$ and standard deviation $\hat{V}_{kj}^{t}$ correspond to:

$$\hat{V}_{kj}^{B} = \hat{\mu}_{j}^{k} - \hat{\mu}^{k} \tag{4}$$

$$\hat{V}_{kj}^{t} = \sqrt{\frac{1}{N^{k}} \sum_{i=1}^{N^{k}} (y_{ij}^{k} - \hat{\mu}^{k})^{2}}. \tag{5}$$

Finally, the global and factor-specific BMI are given by:

$$\widehat{BMI}_{j}^{k} = \frac{\hat{V}_{kj}^{B}}{\hat{V}_{kj}^{t}}$$

$$\widehat{BMI}^{k} = \frac{1}{J} \sum_{j=1}^{J} \widehat{BMI}_{j}^{k}.$$

Due to (2) and (3), the BMI remains invariant to the scale and the shifting of matching scores and is closed in the [-1,1] interval, in which positive values indicate that the factor-specific mean is higher than the overall mean and negative values indicate the opposite. Values of approximately 0 denote the absence of the menagerie effect. To perceive the magnitude of the menagerie effect, Fig. 12 gives the boxplots of the absolute BMI (at left) and BMI values (at right), obtained for the facial expression (left), subject (center) and subject/facial expression joint factors (right), both for genuine (G) and impostor scores (I). Features intrinsic to each subject clearly have much stronger roles in the menagerie effect than facial expressions. When analyzing the joint effect facial expression/subject, the BMI value magnitudes attain maximum values, suggesting that facial expressions stress the menagerie effect for subjects whose intrinsic features make them *a priori* susceptible to that effect. In the plot shown at the right side, note the larger range of both the genuine and impostor bars in the "Joint" group, when compared to the corresponding values in the "Subjects" data series.

Figure 13 shows histograms for the relative frequencies of each facial expression in the first/last decile of the genuine/matching BMI scores, which are used to determine elements of each family in the *menagerie*. The upper-left histogram gives the relative frequencies of each facial expression in the first decile of $\widehat{BMI}^{G}$, meaning that these genuine scores are higher than the global mean and are thus the hardest pairwise comparisons to match (hence considered *goats*). This analysis indicates that *happiness* has a substantially higher probability of being classified as a *goat* than others, whereas *surprise* and *sadness* are most probably in the extremal types of a biometric system (doves and worms), which might be justified by the higher expressivity of most subjects with these expressions. *Chameleons* and *Phantoms* are observed to be complements to each other: *neutral* and *angry* expressions are simultaneously the hardest to match and mismatch, which increases their *chameleon* scores. Globally, *fear* constitutes the most reliable expression for recognition effectiveness, as it receives minimal probabilities in *worms* and *phantoms* and was never among the most probable facial expressions in any category.

Figure 14 illustrates some subjects and corresponding expressions of the most prominent cases in each menagerie family. Near each row, we give the BMI scores that lead to such categorization. The top-left subject has a $\widehat{BMI}^{G}$ genuine score, which is substantially lower than the overall mean for the different facial expressions. The phantom case

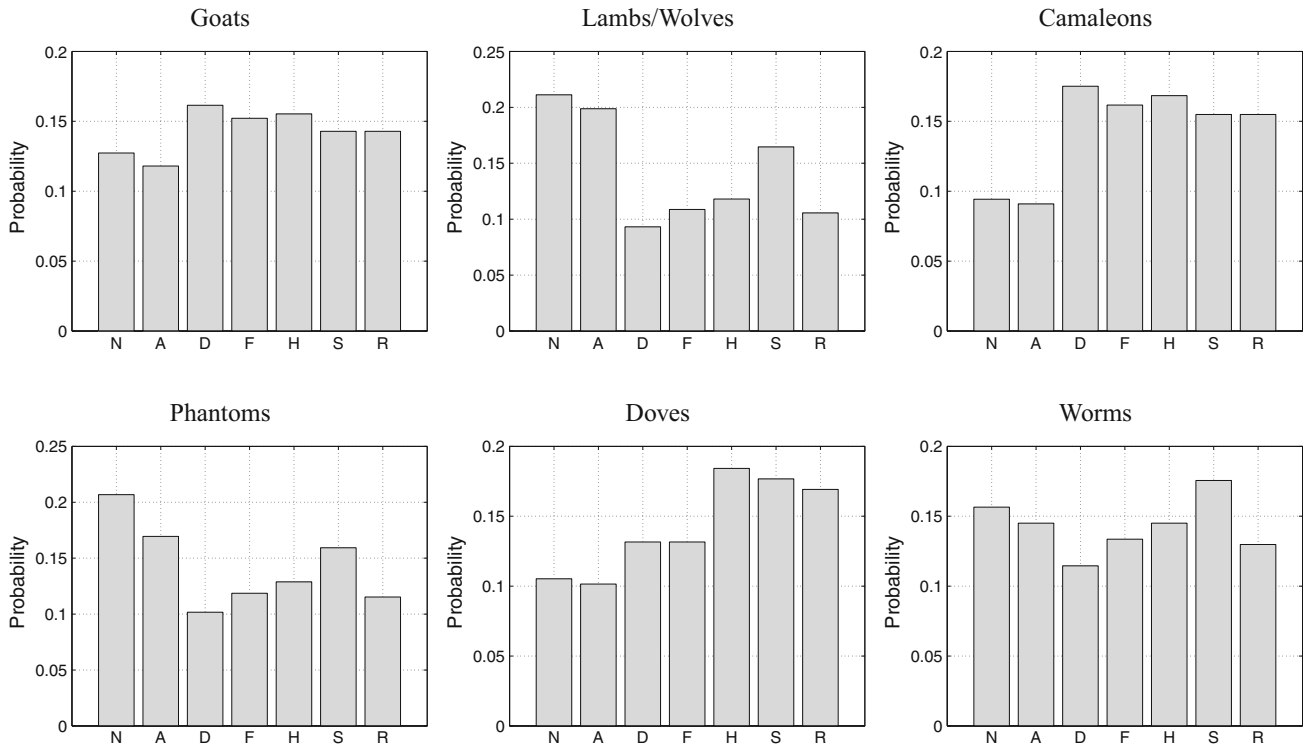


Fig. 13 Prior probabilities of each type of facial expression in the *biometric menagerie* families, based on experiments carried out in the *faceUBIExpress* and *Cohn–Kanade AU-Coded Expression* sets. The symbols *N*, *A*, *D*, *F*, *H*, *S* and *R* denote, respectively, the *neutral*, *anger*, *disgust*, *fear*, *happy*, *sad* and *surprise* facial expressions

is particularly interesting: some females changed their hairstyles between imaging sessions, creating occlusions in the periocular region that were not consistent between sessions and making them difficult to match against themselves and others. Interestingly, when a hairstyle falls in the periocular region and remains unaltered between sessions, subjects tend to be categorized as doves (optimal users), as the hair shape inside the periocular region acts as a discriminating feature. Worms include mostly subjects with evident dynamic changes between frames in their eyelid and eyebrow shapes between expressions, making them vulnerable to impersonalization and difficult to match against themselves.

5 Conclusions

Using the human eye to perform biometric recognition has been gaining popularity, which led to the emergence of the *periocular recognition* field of research. Due to the large number of facial muscles that interact in the periocular region, facial expressions play a significant role in the recognition effectiveness. This paper focuses on such effect, comparing the effectiveness of the most popular periocular recognition strategy for varying facial expressions. Using a dataset in which subjects appear with different facial expressions, we concluded that facial expressions decrease recognition effectiveness in a consistent way, especially when gallery and probe data have different expressions. Such degradation in performance can be reduced if multi-sample gallery data with different facial expressions are available and modules for recognizing facial expressions are enclosed in the recognition process.

Finally, we assessed the role of facial expressions in the well-known *biometric menagerie* effect, having concluded that facial expressions only play a minor role in the categorization of an individual in a menagerie class. However, facial expressions were observed to augment the potential to produce outliers (low) genuine matching scores, which might be particularly concerning for the *goats* and *phantoms* families.

6 Reproducible research

All the information required to reproduce the results given in this paper is available at http://www.di.ubi.pt/~hugomcp/periocularExpressions. This web page contains a link to download all the used images and a set of text files that summarize the most important information:

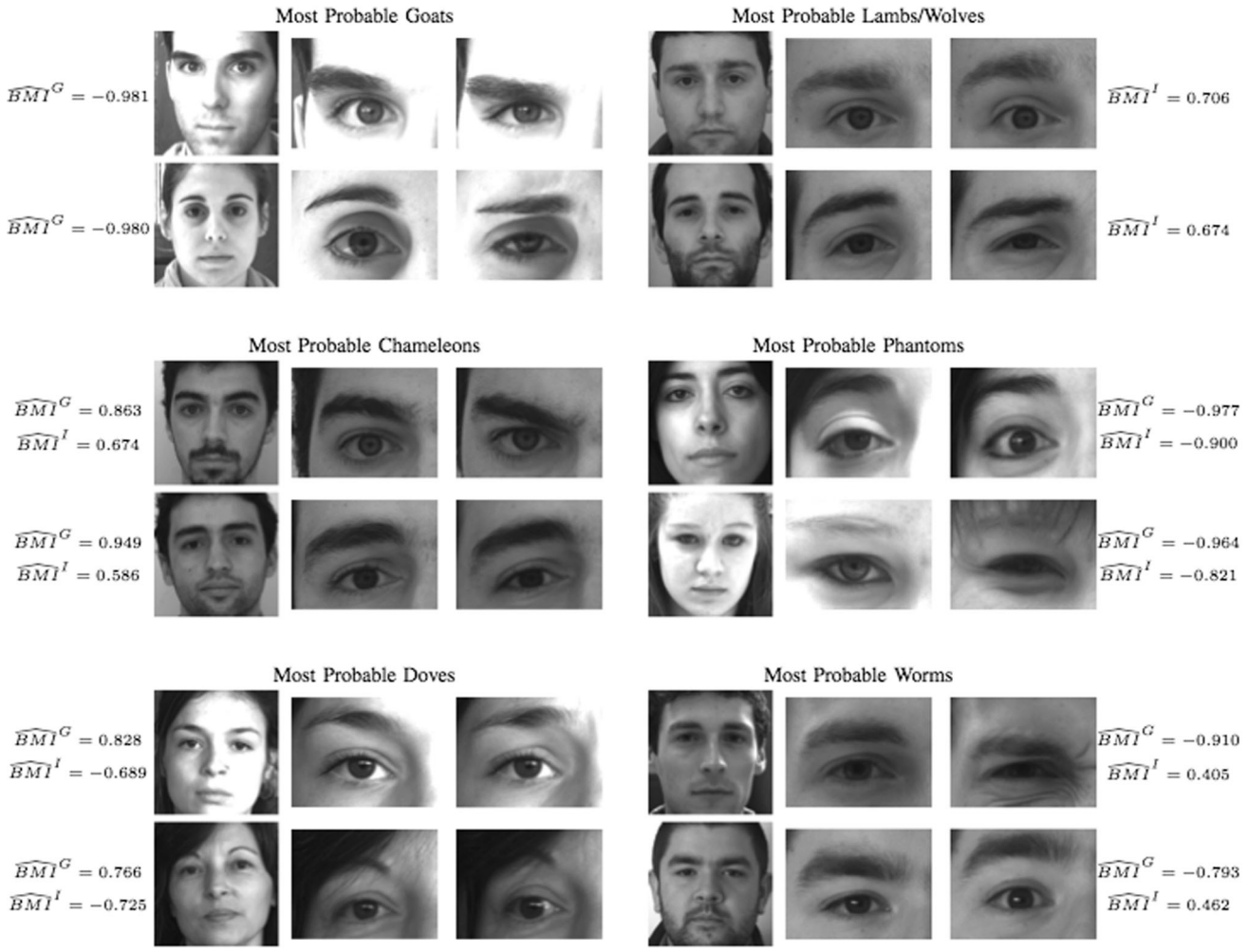


Fig. 14 Examples of the most prominent subjects and corresponding facial expressions regarding the *menagerie* effect. Next to each image, the corresponding BMI indexes ($\widehat{BMI}^G, \widehat{BMI}^I$) stand for the genuine and impostor scores from where the categorization was inferred

- *data.txt* This file contains the description of all the comparisons between pair of images performed, in order to obtain the results given in Sect. 4.
- *ground.txt* This file contains the description about the annotation data that was manually created for each used image.
- *parameters.txt* This file contains the description of the parameters used for every phase that compose the proposed method.
- *packages.txt* This file contains the description of the third party software packages used, and instructions about the way to obtain them.
- *faceExpressUBI.txt* This file contains detailed instructions to access the *FaceExpressUBI* dataset

Acknowledgments This work was carried out in the scope of the research project UID/EEA/50008/2013, R&D Unit 50008, financed by the applicable financial framework (FCT/MEC) through national funds and co-funded by FEDER PT2020 partnership agreement.

References

1. Park U, Ross A, Jain A (2009) Periocular biometrics in the visible spectrum: a feasibility study. In: IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS '09, pp 1–6
2. Park U, Jillela RR, Ross A, Jain AK (2011) Periocular biometrics in the visible spectrum. IEEE Trans Inf Forensics Secur 6(1):96–106
3. Lyle J, Miller P, Pundlik S, Woodard D (2010) Soft biometric classification using periocular region features. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), 2010, pp 1–7
4. Woodard D, Pundlik S, Miller P, Jillela R, Ross A (2010) On the fusion of periocular and iris biometrics in non-ideal imagery. In: 20th International Conference on Pattern Recognition (ICPR), 2010, pp 201–204
5. Bharadwaj S, Bhatt H, Vatsa M, Singh R (2010) Periocular biometrics: when iris recognition fails," in Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), 2010, pp 1–6
6. Park U, Ross A, Jain A (2012) Matching highly non-ideal ocular images: an information fusion approach. In: IEEE 5th International Conference on Biometrics, ICB2012

7. Hollingsworth K, Darnell S, Miller P, Woodard D, Bowyer K, Flynn P (2012) Human and machine performance on periocular biometrics under near-infrared light and visible light. IEEE Trans Inf Forensics Secur 7(2):588–601
8. Woodard D, Pundlik S, Miller P, Lyle J (2011) Appearance-based periocular features in the context of face and non-ideal iris recognition. Signal Image Video Process 5:443–455
9. Crihalmeanu S, Ross A (2011) Multispectral scleral patterns for ocular biometric recognition. Pattern Recognit Lett no. 0
10. Kanade T, Cohn J, Tian YL (2000) Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), pp 46–53
11. Lucey P, Cohn J, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW'10), pp 94–101
12. Ekman P (1999) Facial expressions. In: Dalgleish T, Power M (eds) Handbook of cognition and emotion, John Wiley & Sons, San Francisco, California, USA, pp 301–320
13. Anitha M, Venkatesha K, Adiga BS (2010) A survey of facial expression databases. Int J Eng Sci Technol 2(10):5158–5174
14. Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A (2010) Presentation and validation of the Radboud Faces Database. Cognit Emot 24(8):1377–1388
15. Ebner C, Riediger M, Lindenberger U (2010) FACES-a database of facial expressions in young, middle-aged, and older women and men: development and validation. Behav Res Methods 42(1):351–62
16. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with Gabor wavelets. In: Third IEEE International Conference on Automatic Face Gesture Recognition. IEEE Computer Society, Nara, Japan, pp 200–205
17. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, pp 317–321
18. Bettadapura VK (2009) Face expression recognition and analysis : the state of the art. Emotion, pp 1–27
19. Haq S, Jackson P (2010) Machine audition: principles, algorithms and systems. In: Multimodal Emotion Recognition. IGI Global ch., Hershey PA, pp 398–423
20. Sebe N, Lew M, Sun Y, Cohen I, Gevers T, Huang T (2007) Authentic facial expression analysis. Image Vis Comput 25(12):1856–1863
21. Sim T, Baker S, Bsat M (2003) The cmu pose, illumination, and expression database. IEEE Trans Pattern Anal Mach Intell 25:1615–1618
22. Gross R (2005) Face databases. In: Handbook of face recognition. Springer, ch 13, pp 301–327
23. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The feret evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104
24. Hwang BW, Byun H, Roh MC, Lee SW (2003) Performance evaluation of face recognition algorithms on the asian face database, kfdb. In: Proceedings of the 4th international conference on Audio- and video-based biometric person authentication, ser. AVBPA'03. Springer-Verlag, Berlin, Heidelberg, pp 557–565
25. O'Toole AJ, Harms J, Snow SL, Hurst DR, Pappas MR, Ayyad JH, Abdi H (2005) A video database of moving faces and people. IEEE Trans Pattern Anal Mach Intell 27:812–816
26. Yin L, Wei X, Sun Y, Wang J, Rosato MJ (2006) A 3D facial expression database for facial behaviour research. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06). IEEE Computer Society
27. Guyon I, Makhoul J, Schwartz R, Vapnik V (1998) What size test set gives good error rate estimates ? IEEE Trans Pattern Anal Mach Intell 20(1):52–64
28. Cantor ABM (2002) Understanding logistic regression. Evid Oncol 3(2):52–53
29. Yager N, Dunstone T (2010) The biometric menagerie. IEEE Trans Pattern Anal Mach Intell 32(2):220–230
30. Poh N, Kittler J (2009) A biometric menagerie index for characterizing template/model-specific variation. In: Proceedings of the Third International Conference on Advances in Biometrics-BTAS 09, pp 816–827