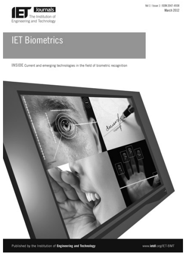
Published in IET Biometrics
Received on 15th July 2014
Revised on 17th September 2014
Accepted on 23rd September 2014
doi: 10.1049/iet-bmt.2014.0045



ISSN 2047-4938

BioHDD: a dataset for studying biometric identification on heavily degraded data

Gil Santos[1], Paulo T. Fiadeiro[2], Hugo Proença[1]

[1]Department of Computer Science, IT – Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
[2]Department of Physics, Remote Sensing Unit – Optics, Optometry and Vision Sciences Group, University of Beira Interior, Covilhã, Portugal
E-mail: gmelfe@ubi.pt

Abstract: Substantial efforts have been put into bridging the gap between biometrics and visual surveillance, in order to develop automata able to recognise human beings 'in the wild'. This study focuses on biometric recognition in extremely degraded data, and its main contributions are three-fold: (1) announce the availability of an annotated dataset that contains high quality mugshots of 101 subjects, and large sets of probes degraded extremely by 10 different noise factors; (2) report the results of a mimicked watchlist identification scheme: an online survey was conducted, where participants were asked to perform positive and negative identification of probes against the enrolled identities. Along with their answers, volunteers had to provide the major reasons that sustained their responses, which enabled the authors to perceive the kind of features that are most frequently associated with successful/failed human identification processes. As main conclusions, the authors observed that humans rely greatly on shape information and holistic features. Otherwise, colour and texture-based features are almost disregarded by humans; (3) finally, the authors give evidence that the positive human identification on such extremely degraded data might be unreliable, whereas negative identification might constitute an interesting alternative for such cases.

1 Introduction

The evolution of the concept of biometrics over the last decades is linked with societies' increasing concerns about both individual and global security. From personal computers to border access control everyone aims at securing their identities, their assets and, primarily, their homeland. Such safety relies on the ability to accurately identify subjects based on biometric features, either biological or behavioural.

Biometric systems rely on the accurate 'extraction' of individuals' distinctive features and their proper 'encoding', so that the essential information is preserved. Those requisites are traditionally assured by high acquisition constraints, with the subject cooperation being a key-element. When moving to unconstrained scenarios, those acquisition constraints are lowered and subject cooperation is not expectable. Recognition became more challenging and alternatives are sought [1, 2], either by: (1) improving the existing algorithms; (2) resorting to multi-modal biometric systems; or (3) exploring new traits could better cope with this new reality. Despite those efforts, no system yet exists capable of effectively dealing with all the issues introduced by biometrics 'in the wild'. In fact, even biometric systems able to cope with less constrained conditions (e.g. Iris-on-The-Move project [3]) still lack an ideal level of user abstraction.

Visual surveillance is a very active field in computer vision, with a lot more applications other than biometrics 'per se' [4]. Existing automatic surveillance systems are rather focused on activity recognition (e.g. W^4 project [5]), and not many projects are prepared to deal with surveillance scenarios from a watchlist approach (e.g. Kamgar-Parsi *et al.* [6]). Furthermore, none of the latter works from the negative identification perspective.

Most biometric systems attempt positive identification (or verification) against a gallery of enrolled users based on a (dis)similarity measure. In many 'in the wild' applications however, biometric systems make more sense when used from the negative perspective: guarantee with enough confidence that an unknown subject does not belong to a gallery of 'persons-of-interest', instead of attempting to identify him. On that basis, facing a watchlist scenario one can aim at spotting a distinctive feature on the probe subject, and exclude those who neither share that feature, nor any of its possible transformations. Moreover, even if we do not have enough distinctive features to support a positive recognition (e.g. because of the quality of acquired images) we can still perform reliable negative recognition.

1.1 Contextualisation: facial biometrics

The everyday use of facial cues includes recognising our peers or unveiling their state of mind, which happens seamlessly and unawarely. Is then easy to place face as the most common and widely used biometric trait, and one of the most successful applications of image analysis and understanding. Several face recognition systems are commercially deployed and a lot of techniques accessible [7], working on both still and video images. Algorithms are based either on the global analysis of the whole image, or on the relation between facial elements, their localisation and shape. In either case, their effectiveness is conditioned by several factors, which become even more evident 'in the wild': its three-dimensional structure lead to substantial differences in appearance, accordingly to the subject's pose; large portions are often occluded on non-orthogonal data acquisition; facial expressions affect their appearance; and its particularly easy to disguise.

Analysing the human ability to recognise each other, researchers can identify the more reliable cues, valuable for the develop well-grounded recognition methods. Previous studies report interesting findings when exploring the human ability to identify faces (e.g. Sinha *et al.* [8]), encouraging further researching on understand how people cope with 'in the wild' circumstances. In this study we do not aim at mimicking the identification process taking place in human vision, but rather to provide useful insights for further research on this topic. We analyse the noise factors' impact on human identification performance, identifying the features people recall as basis for their judgement.

1.2 Contextualisation: similar datasets

Publicly available datasets exist for both video surveillance [4] and face biometrics [9] research, acquired under less constrained conditions. Although a much higher extent of databases is available, five significant datasets must be mentioned, which contain a more significant amount of pie changes: FERET [10], CMU-PIE [11], CAS-PEAL [12], Multi-PIE [13] and LFW [14] (Table 1). The latter two datasets are presumably the most completes, each one by its own reasons: the Multi-PIE provides facial images from 337 subjects, imaged over four sessions under 15 pose and 19 illumination variations, along with high-resolution registration photos; the Labeled Faces in the Wild (LFW) dataset contains a larger amount of images and subjects, 13 233 and 5749, respectively, at completely 'in-the-wild' conditions, and thus without uniformity among subjects. Although not being an extensive listing of the existing datasets, the ones we present are the most directly comparable to the one we are now establishing.

In this paper we introduce a newly created dataset of heavily degraded facial images, where the 'in the wild' conditions associated with visual surveillance systems are closely simulated. Full 360° illumination and pose variations are introduced, among with other realistic noise factors at different reasonable levels, along with ground-truth information for research validation. Despite containing a lower amount of participants when compared to the existing databases, this new dataset contains a wider range of pose and illumination variations, uniform and comparable for all subjects.

Table 1 Overview of the most relevant and public available face recognition datasets with pie variations, with comparison to our working dataset

Dataset	Subjects	Sessions	Pose	Illumination	Expression
FERET	1199	2	20	2	2
CMU-PIE	68	1	13	43	4
CAS-PEAL	1040	2	21	15	6
multi-PIE	337	4	15	19	6
LFW	5749	?	?	?	?
BioHDD	101	2	24	72	1

Values marked with '?' can not be determined because of the nature of the dataset

The remainder of this paper is organised as follows: Section 2 describes the BioHDD dataset, detailing the acquisition framework, enrolled participants and introduced noise factors; Section 3 presents the experimental method used in our study, with a thorough analysis of its results; finally, Section 4 states some final considerations, along with further lines of work.

2 BioHDD dataset

The main objective of the BioHDD database was to gather images from a significant group of individuals, ranging from clear frontal shots to heavily degraded facial images, enabling to assess the feasibility of biometric recognition 'in the wild'.

2.1 Imaging framework and setup

The imaging framework was installed in a closed lounge without uncontrolled lightening sources. Participants were illuminated with a single 800 W halogen projector, and a white cloth was used as image background to avoid contextual interferences. The acquisition process consisted of three acquisition stages: registration, still image acquisition and video acquisition.

At the registration stage three reference facial images of high-quality were acquired from each participant (frontal, left- and right-hand side – Fig. 1). The acquisition device gathered information from the visible wavelength slice of the electromagnetic spectrum, with the light source directly above it. Participants were asked to essay a neutral expression and look forward, aided by three fixation points, so that all observers were facing the same direction during this stage.

On a second stage images were acquired 'simultaneously' on both NIR and VIS, while introducing four variations: illumination angle and intensity, subject revolution and head-tilt – Fig. 3, columns 1–4. Changes on the illumination angle were achieved with the halogen projector shifting on 45° steps (Fig. 2, *A*–*H*), while participants kept facing the acquisition device. Additionally, participants were asked to face eight fixation points evenly distant from each other, introducing subject revolutions in full 360°. For all variations, participants were imaged facing forward and tilting their head up and down, while simulating illumination intensity changes using the acquisition device exposure settings.

At a final stage, subjects walked trough a corridor with non-uniform illumination conditions whilst captured by a VIS greyscale camera placed on a upper level. As we can

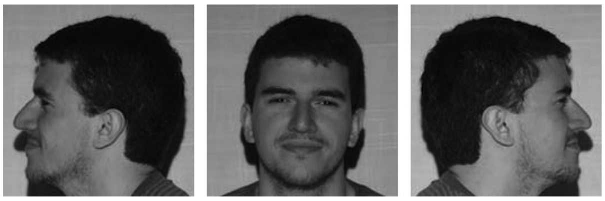


Fig. 1 *Example of images acquired used as gallery data: left-hand side, frontal and right mugshots*

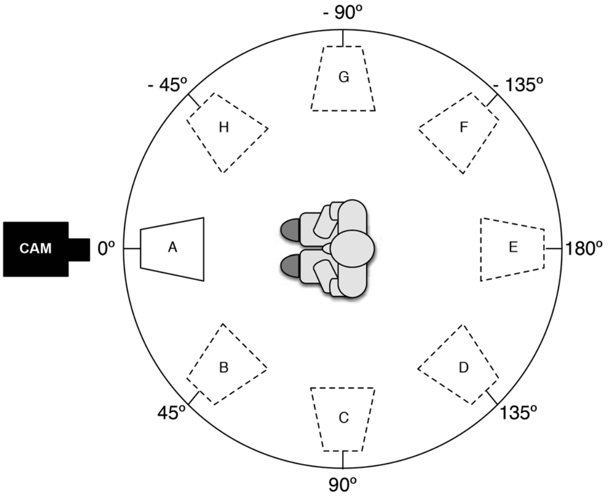


Fig. 2 *Schematic perspective of the image acquisition framework (over-top view)*

For illumination changes, the light source alternate on positions A to H with participants facing the camera
For rotation changes, participants were asked to align themselves with the different reference points while the camera and light source remained aligned at the initial position

see from the samples at Fig. 4, surveillance-like data acquisition was closely simulated.

Table 2 presents a complete hardware specification.

Data was gathered on two acquisition sessions with a minimum of two weeks apart. On the first acquisition session participants wearing glasses were required to remove them, and the ones with longer hair were asked to tie it. Likewise, videos acquired during that same session had participants looking at a fixation point while walking. To increase variability, on the second session such constraints were not applied. No modifications were introduced on the hardware setup or location. Attendance to both sessions was around 88%, representing a total of 101 participants. As described on Table 3, 66 male subjects and 35 female subjects were enrolled, most of them Caucasian. For normalisation purposes, acquired images were manually cropped to 600 × 600 px, while keeping the face centred. Registration images from Stage 1 were stored with 2, 048 × 2, 048 px.

2.2 Heavily degraded data

Not all noise factors associated with recognition ‘in the wild’ were introduced during the acquisition stage. As so, additional image degradation procedures were carried on.

A total of ten noise factors were identified and grouped in three different sets: (1) ‘real’ noise factors introduced with the imaging setup; (2) ‘simulated’ noise factors that although not introduced at the acquisition stage are related with the imaging process; and (3) noise factors associated with data storage and transmission. Each noise agent comprises different levels (L_i), as illustrated on Fig. 3, and their presence follows the reasoning we now describe.

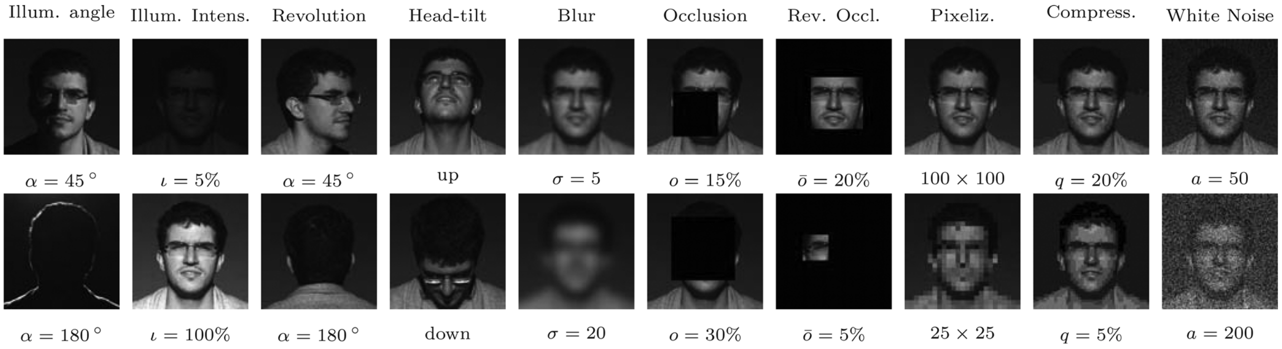


Fig. 3 *Examples of the types of image degradation factors in the BioHDD dataset*

From left- to right-hand side: illumination angle and intensity, subject revolution, head-tilt, blur, occlusion and reverse occlusion, pixelisation, compression and white noise
The top row corresponds to the first noise level L_1, and the bottom row to the maximum noise level $L_{\max}$
On illumination intensity and head-tilt, both images represent L_1, since their difficulty is similar
Although only VIS data is depicted, each image has its NIR counterpart

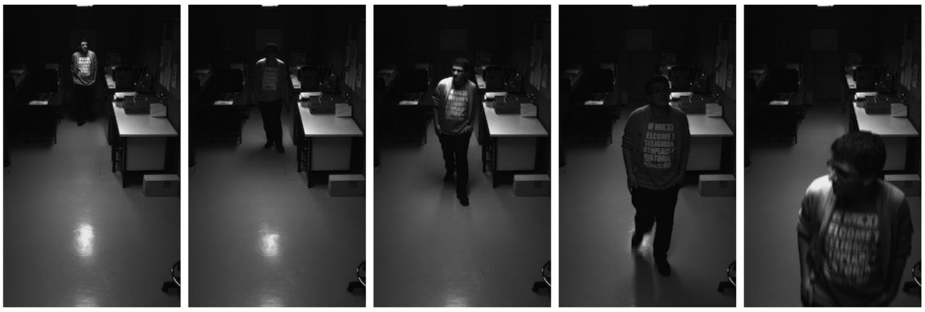


Fig. 4 *Samples from the video acquisition stage*

Frames were cropped for illustration purposes

Table 2 Details of the BioHDD acquisition devices, image and video settings

	Registration	Image Acq.	Video Acq.
camera	Canon EOS 5D	JAI AD-080GE	Stingray F504-B
lens	Canon EF 100-400	NIKKOR 55-80	HR F1.4/8 mm
spectrum	visible	visible + NIR	visible
color space	RGB	RGB + greyscale	greyscale
channel depth	8bit	8bit	8bit
frame size	4368 × 2912 px	984 × 768 px	1224 × 1028 px
cropped size	2048 × 2048 px	600 × 600 px	–
format	PNG	PNG	AVI
frame-rate	–	–	15 fps

2.2.1 Real noise factors: As previously described, this set of noises was directly introduced at the acquisition stage. When working in unconstrained scenarios optimal illumination cannot be assumed. Along with the images captured at the 'best' conditions (with average exposure and having the light source directly over the acquisition device), data was also captured varying the 'lightening angles' and the 'illumination intensities' (low lighting and over-exposure). The chosen angles cover all 360° degrees (at 45, ° steps), and intensity changed from 5 to 100%. To cover a higher amount of poses, subject 'revolution' was also introduced over eight angles (similarly to illumination) and 'head-tilting' in two, with participants facing up and down. Those choices were based on the reasoning that individuals trying to avoid detection are most likely to be facing the ground or away from any visible cameras.

2.2.2 Simulated noise factors: To mimic acquisition issues as the ones associated with inappropriate lens settings, poor focus, subject movement etc, four levels of 'blur' were simulated applying Gaussian filters with standard deviation raging from $\sigma_{L_1} = 5$ to $\sigma_{L_4} = 20$.

Face occlusion was simulated by overlapping a black patch to the original image, covering $o_{L_1} = 15\%$ to $o_{L_1} = 30\%$. A different flavour of occlusion where only a small portion of the image is left-hand side visible, $\bar{o}_{L_1} = 20\%$ to $\bar{o}_{L_4} = 5\%$, was also simulated. This noise factor can also be related with the use of certain headgear (e.g. balaclava).

In certain scenarios we observe that the used devices are of low or insufficient spatial resolution, or post-processing censorship is applied to avoid detection of a particular subject or distinctive feature that is intended to remain anonymous. This 'pixelisation' effect was obtained by downscaling the original photo: $\text{size}_{L_1} = 100 \times 100\,\text{px}$ to $\text{size}_{L_4} = 25 \times 25\,\text{px}$.

2.2.3 Storage/transmission related noise factors: Finally, 'compression' degradation found on systems that rely on digital storage or broadcasting was simulated using a standard JPEG algorithm. Quality ranged from $q_{L_1} = 20\%$ to $q_{L_4} = 5\%$. Based on the same reasoning, inherent to data storage on photographic film or broadcasted through analogue channels, 'white noise' was simulated.

To generate probe images I_P, one transformation from each set T_1, T_2, T_3 was randomly selected, and the corresponding noise factor was applied to the original image at a random level k, l, m, respectively (1). Noise application was sequential, with the last noise transformation T_3 being applied upon T_2 result, denoted by ° and T_2 being applied over T_1 output. Sample probe images obtained using this fusing technique are illustrated at Fig. 5

Table 3 Details of the BioHDD subjects that offered themselves as volunteers to both imaging sessions

Gender	Male	65%	Age	[0, 20]	10.89%
	Female	35%		[21, 25]	44.55%
Origins	European	95%		[26, 30]	15.84%
	African	4%		[31, 35]	9.90%
	Asian	1%		[35, 99]	18.81%

$$\begin{aligned} I_P &= (T_3(L_m) \circ T_2(L_l) \circ T_1(L_k))(I) \\ &= T_3(L_m)(T_2(L_l)(T_1(L_k)(I))) \end{aligned} \qquad (1)$$

2.3 Dataset availability

The complete BioHDD dataset is public and freely available for academic and research purposes [http://biohdd.di.ubi.pt]. Researchers are granted access to: (1) 606 registration images; (2) 27 270 probe samples with the variations introduced during the acquisitions stage; (3) 27 270 similar images on the NIR spectrum; (4) 2500 probe images with combined noises; (5) 202 greyscale videos with surveillance like data from each participant. Further probe images can be generated 'on-demand', and all data comes with ground-truth information about the associated noise levels.

3 Experiments and discussion

3.1 Experimental method

Our goals to study the human ability to identify their peers on heavily degraded data were: (1) identify the noise factors whose avoidance would be preferable, by associating each one of them with a specific impact on human identification performance; (2) pin down the regions identified as part of the process and, if possible, even specific features; (3) illustrate how negative recognition might still be reliable 'in the wild', where the positive approach is unattainable. To do so, a web-page was built with a custom participation interface mimicking a watchlist recognition scenario – Fig. 6.

For this experiment, a total of 200 000 trials were generated, combining 2500 probe images and 2500 galleries. At the begging of each test the interface was populated with a random trial, with 3/4 probability of the gallery containing the subject on the probe image. Each participant was asked to do one of three actions, for each gallery identity shown

1. mark it as green if they feel that the identity on the mugshot corresponds to the probe image (positive identification);
2. mark it red if they are certain that the identity on the mugshot does not correspond to the query image (negative identification);
3. leave it blank, in case of uncertainty.

In the case of identification, participants were asked to fill the appropriate text-box to justify their answer. No time restriction was set for image examination, and upon finishing a new test was loaded. Each participant was free to take as many tests as he wanted. The experiment ended after one month, collecting a total of 3650 participations from 45 different countries. A total of 17 438 identifications and 1422 justifications were obtained.

doi: 10.1049/iet-bmt.2014.0045

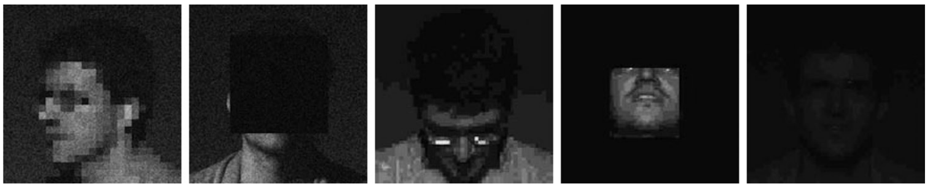


Fig. 5 *Sample trial images with different levels of noise combined*

3.2 Results and discussion

Although including a third class for ‘no decision’ in our testing interface, we simplified our problem to a binary one by analysing the answers where participants were sure enough of their answer to give a specific identification (either positive or negative). To assess the identification performance, four well-known statistical measures were used: sensitivity (or true positive rate [TPR]), specificity (SPC), accuracy (ACC) and Matthews correlation coefficient (MCC). TPR and SPC, are given by (2) and (3), respectively, and weight the correct responses by the total of positive (true positives (TP) + false negatives (FN)) and negative (true negatives (TN) + false positives (FP)) answers

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

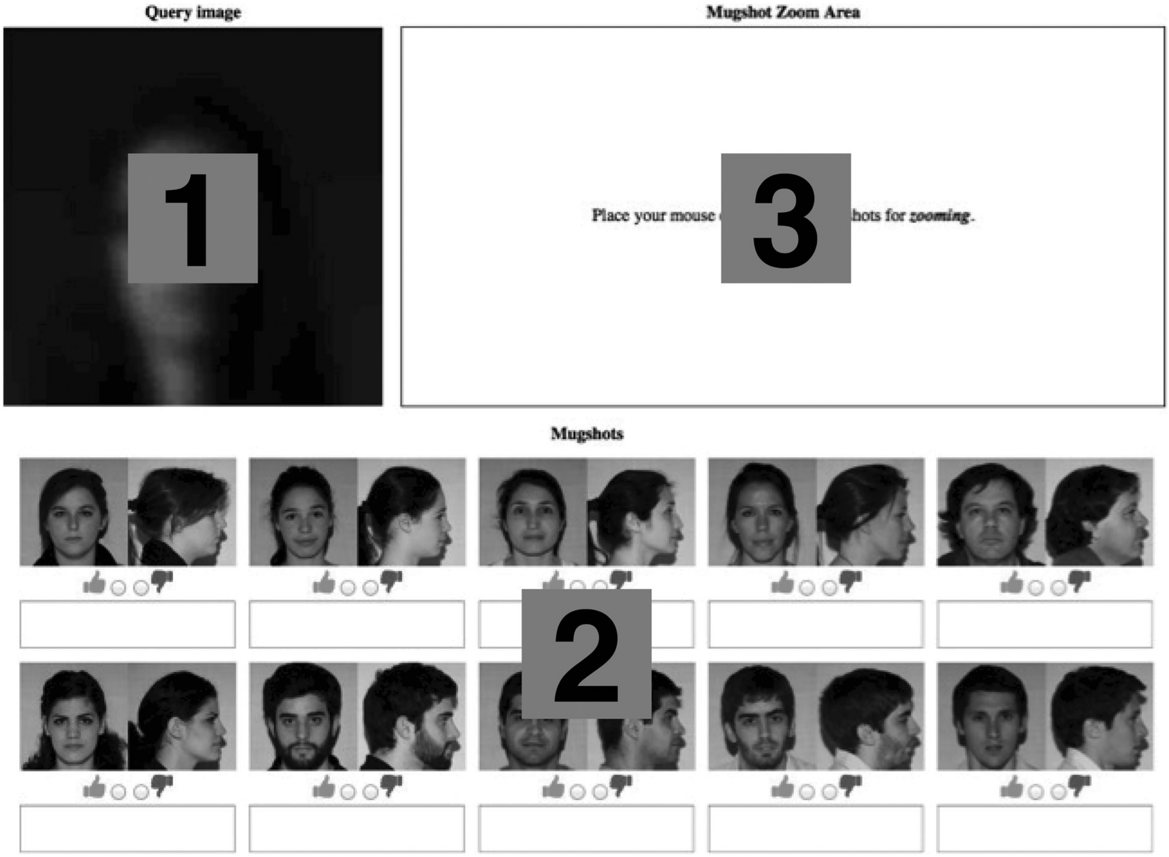
$$\mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

The accuracy gives us the overall ratio of correctly classified matches, 1 being the optimal value where all instances have been correctly classified. For a balanced analysis we used MCC, which takes into account the high discrepancy between the amount of positive and negative matches. It can be regarded as a correlation coefficient between participants’ answers and the correct identification, and its output ranges in the $[-1, 1]$ interval, where 1 the optimal value [15]

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{4}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{5}$$

In Fig. 7 we can see from the true positive rate and specificity probability density functions, computed for all subjects on the dataset. TPR is clearly more prone to variations, as positive samples are more difficult to be found in the experimental



Place your mouse hots for *zooming*.

Fig. 6 *Web interface of the conducted survey, with three major panels: (1) a probe sample from an unknown identity; (2) a set of 10 profile / frontal mugshots, representing the gallery dataset; (3) zoomed-up perspective of each gallery sample, populated on mouse-over on region 2*

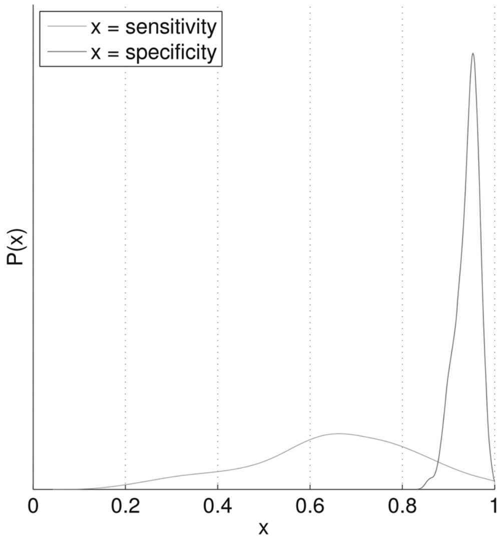


Fig. 7 *Per-subject sensitivity and specificity probability density functions*

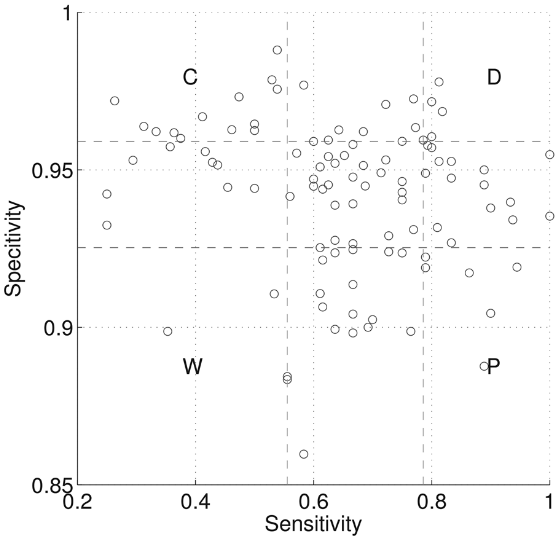


Fig. 8 *Zoo plot for the overall user performance*

Dashed lines represent the first and third quartiles for sensitivity and specificity distributions
The identities on the 'P','D', 'C' and 'W' regions are more likely to assume dove (D), chameleon (C), phantom (P) and worm-like (W) behaviour

setup. We should have in mind that one out of four trials could not lead to TP, since the subject from the probe image is not in the gallery. To stress the possible relationship between false positives and the impossibility of making a positive match, implying participants had attempted identification either way, we performed a paired-sample Student's t-test: For all the n subjects in the database, let us define the fall-out distributions d_1 for trials where positive matches were possible, and d_2 when not. Let us then consider the null hypothesis H_0 where the difference D between d_1 and d_2 follows a normal distribution with mean equal to zero and unknown variance, tested through (6)

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}} \tag{6}$$

where $\overline{X}_D$ and s_D are D average and standard deviation values, and μ_0 the mean for the d_1 distribution. Experimental data returned a $p$value of 1.81×10^{-14}, thus rejecting the null hypothesis: the distributions are significantly different, indicating that most participants indeed attempted to perform a positive match, even when it was not possible.

Plotting each one of the subjects in the dataset as a function of the TPR and SPC, we can understand their individual propensity to correct identification – Fig. 8. Furthermore, we can group them in four biometric menagerie classes as suggested by Yager & Dunstone [16]: doves, chameleons, phantoms and worms. 'Doves' are the most favourable subjects and the optimal group for any recognition system, as they do not produce verification error. High values are observed for both TPR and SPC. 'Chameleons' are subjects who are easily misidentified as they always appear similar to others, their specificity is high, but true positive rate is extremely low. 'Phantoms', in opposition to chameleons, are associated with low SPC and high TPR. 'Worms', contrary to doves, are the most critical subjects in a biometric system. They behave in the worst possible way, yielding low true positive rate and specificity. At a central location we have 'the herd', where the most common users ('sheep') are located.

To define the limits for each class, we start by defining two regions for the true positive rate, TPR_{Q1} and TPR_{Q3}, containing the subjects below the first quartile and over the third quartile, respectively. If we define two similar regions (SPC_{Q1} and SPC_{Q3}) for the specificity, a subject s is said to assume a particular behaviour according to (7) [17]

$$\begin{cases} \text{Dove}, & \text{if } s \subset \text{TPR}_{Q3} \cap \text{SPC}_{Q3} \\ \text{Chameleon}, & \text{if } s \subset \text{TPR}_{Q1} \cap \text{SPC}_{Q3} \\ \text{Phantom}, & \text{if } s \subset \text{TPR}_{Q3} \cap \text{SPC}_{Q1} \\ \text{Worm}, & \text{if } s \subset \text{TPR}_{Q1} \cap \text{SPC}_{Q1} \end{cases} \tag{7}$$

As we can see on Table 4, even with the degradation introduced in the probe images participants were able to correctly match 92% of the instances they were presented with. To assess the effect of each noise on that performance level, we computed the same metrics for when removing each one of them. Additionally, we analysed how each menagerie class relocated as a consequence of a specific noise, as follows.

Take an initial point $A(\text{TPR}_a, \text{SCP}_a)$ representing the global recognition capabilities of an individual on the dataset, and a point $B(\text{TPR}_b, \text{SPC}_b)$ computed likewise for when a specific noise is removed. We can then compute the global distance to the optimal point $O(1, 1)$ as (2), and the distance upon noise removal d_b likewise.

$$d_a = \sqrt{(1 - \text{TPR}_a)^2 + (1 - \text{SPC}_a)^2} \tag{8}$$

Finally, the individual optimisation produced by noise removal can be accessed through $\zeta \to [-1, 1]$, where -1 represents the worst case scenario and 1 the best improvement possible. Zero means no performance change

$$\zeta = \frac{d_a - d_b}{d_a + d_b} \tag{9}$$

Assessing the average $\zeta -$ values on each one of the zoo-plot regions, we obtain the values at Table 5.

doi: 10.1049/iet-bmt.2014.0045

Table 4 Overall sensitivity (TPR), specificity (SPC), accuracy (ACC) and MCC values and the same statistics for when a noise factor is removed

	TPR	SPC	ACC	MCC
overall	0.657	0.941	0.918	0.547
illumination angle	0.682	0.944	0.922	0.573
illumination intensity	0.633	0.938	0.913	0.518
revolution	0.641	0.941	0.916	0.537
head-tilting	0.671	0.941	0.919	0.558
Gaussian Blur	0.670	0.943	0.920	0.560
occlusion	0.641	0.939	0.914	0.532
rev. occlusion	0.675	0.942	0.920	0.558
pixelisation	0.641	0.941	0.917	0.537
compression	0.618	0.937	0.911	0.506
white noise	0.688	0.945	0.923	0.580

When a noise factor is withdrawn, one would expect the optimisation to always be positive. However, both analyses show only four noise factors that led to significant improvements. The most conditioning element is the introduction of the 'white noise' associated with analogue channels, and as the opposite tendency is observed for digital 'compression', we can conclude that digital channels should be used. The second considerable constraint is the 'illumination angle': when the subject being identified was not frontally lit, participants exhibited higher error rates. On the other side, variations on the 'lightening levels' were not relevant, as participants were able to accommodate to both under- and over-exposure. We also observed their ability to cope with 'occlusion' up to a certain degree, and only when a portion of the face was visible ('reverse occlusion') their performance started to degrade. Finally, participants' performance was also significantly conditioned by 'head-tilting'. This last observation is of special importance: as mentioned before, individuals trying to avoid detection are most likely to be facing the ground or away from any visible cameras. Along with illumination intensity and occlusion, some other noise factors' removal did not led to improvements in performance: 'revolution', indicating that useful features can also be derived from the side of the head, and are actively used in human identification; and 'pixelisation', that along with 'compression' lead us to infer that global features are preferred over local and more detailed ones.

When performing a 'per' species analysis, its perceptible how sensitivity tens to decrease at an higher rate than specificity increases. That explains the negative ζ-values for species located over the TPR' third quartile (doves and phantoms), associated with a convergence to 'the herd'. The class that benefits the most from noise removal is 'Worms', with improvements over three times greater than those observed for 'Sheep'.

Table 6 Probability (%), sensitivity TPR, specificity (SPC), accuracy (ACC) and MCC values for feature category usage on recognition justifications

	(%)	TPR	SPC	ACC	MCC
shape	49.64	0.87	0.88	0.88	0.62
color	6.05	0.71	0.92	0.90	0.52
texture	0.51	1.00	0.50	0.71	0.55

As above stated, a set of justifications for each of the responses given by the volunteers of our on-line survey we're collected. These answers are an important source of information to perceive the type of features predominantly used by humans in identification tasks, as well as to relate the usability of each feature to the degree of success in the corresponding identification. Hence, the responses were grouped by the type of feature they mention and the facial region, as detailed in Tables 6 and 7.

On a 'per' category analysis (Table 6), we can see how almost half the justifications mention shape related features, making it the most commonly used feature type. Colour related features are much less used (6.04%), skin and hair colour being the most significant ones. This is a considerable difference, even considering that the dataset consists mainly of young European participants. Attending to the accuracy levels alone, one could be biased into considering the latter to be a better feature.

To take into account both the high specificity value and the difference in class sizes MCC was also analysed. This measure weights the importance of TPR and SPC by the size of each class, shows shape to be not only the most used feature type, but also the more reliable on both positive and negative identification. Finally, the number of participants that used textural information is almost residual (0.51), and usually refers to freckles and another skin signs, tattoos and jewellery.

In Table 7 we summarise the 'per' region analysis. As we can see, when looking to justify the identifications they make participants use holistic features on almost 2/3 of the justifications, with two most relevant cues: perception on probe subject gender, and a broad analysis of head's shape. From that, special attention is paid to top regions, which can intuitively be related to a higher amount of detail, as more elements are present. Actually, if we analyse the weighted accuracy average per region we can see how topmost areas are indeed less deceiving than lower ones, which is explained by the high volume of negative identifications based on hairstyle. Hair related features played an important

Table 5 Average ζ-values for all zoo-plot regions ($\times 10^{-2}$) upon noise removal

	Doves	Chamel.	Phant.	Worms	Sheep	Average
illum. angle	6.66	3.78	−3.26	11.90	3.01	2.79
illum. intens.	4.48	−5.32	−6.19	−8.89	−6.82	−3.00
revolution	−17.27	−0.37	8.47	1.23	−2.37	−2.51
head-tilting	4.58	0.70	−0.61	−0.13	4.64	2.17
Gaussian Blur	−0.81	−0.88	4.26	−1.79	5.76	0.92
occlusion	5.26	−0.96	−6.41	−3.02	−6.73	−1.06
rev. occlusion	1.84	0.82	2.90	3.84	4.91	2.69
pixelisation	−13.49	−0.03	−12.04	7.86	−3.93	−4.38
compression	13.61	−0.83	−19.47	−14.57	−3.99	−3.98
white noise	−17.74	5.74	10.99	13.49	2.71	3.33
average	−1.29	0.27	−2.14	0.99	0.28	

Table 7 Probability (%), sensitivity (TPR), specificity (SPC), accuracy (ACC) and MCC values for feature usage as recognition justification

	(%)		(%)	TPR	SPC	ACC	MCC
holistic	64.29	gender	35.86	0.50	0.99	0.98	0.38
		age	2.33	0.75	0.96	0.94	0.71
		face/head	22.89	0.86	0.93	0.92	0.68
		skin	3.21	0.60	0.90	0.86	0.43
upper face	25.73	hair	23.03	0.74	0.89	0.87	0.54
		forehead	2.70	0.86	0.93	0.92	0.75
mid face	24.28	eyebrows	4.66	0.85	0.78	0.80	0.53
		eyes	3.94	0.91	0.79	0.81	0.59
		glasses	2.26	0.33	1.00	0.94	0.56
		ears	3.43	0.93	0.67	0.74	0.54
		nose	9.99	0.96	0.85	0.88	0.71
lower face	16.11	cheeks	0.36	1.00	0.50	0.60	0.41
		beard/mustache	4.52	1.00	0.89	0.90	0.64
		mouth	2.26	0.89	0.82	0.84	0.66
		chin/jaw	6.05	0.86	0.88	0.88	0.53
		neck	2.92	0.80	0.71	0.73	0.36
other	2.33	shoulders	0.15	1.00	0.00	0.50	—
		clothes	2.19	0.50	0.86	0.83	0.24

Features are grouped per type/region and sub-region.

role as decision factor, being mentioned in almost 1/4 of the answers. Allusions to the forehead were also mostly related to hair-to-skin boundaries, and if we group them as 'upper face' we cover 25.73% of the answers. The second most used region was 'mid-face', whose observation aided on justifying 24.28% of the identifications. Here, periocular information was the most used (10.86%), closely followed by the nose information. From the lower face, the most mentioned feature is a mix of the chin/jaw shape and the texture (the presence of facial hair).

When balancing positive and negative identifications through MCC, we can see how the mid-face is the less deceiving area. For the holistic features, age was the most effective recognition factor. As most of the database participants are young adults (academic students), the ones older than them (academic staff) are easily spotted.

3.3 Positive against negative identification

The degree to which we can rely on positive identification changes significantly when the decision environment degrades. To illustrate a poor decision environment, we computed entropy η as the single feature for subject identification over images acquired at the first three levels of illumination angle and subject revolution. Let I be an image in this set and x_i a pixel intensity level on the [0, 255] interval. Using histogram counts to estimate its relative frequency $P(x_i)$, the global image entropy is given by (10). Attending to probability densities (Fig. 9), we can verify how that constitutes a poor decision environment for any Bayesian classifier to perform positive identification, as functions overlap

$$\eta(I) = \sum_i P(x_i) \log_2 P(x_i) \tag{10}$$

Yet, assuming a null hypothesis H_0 corresponding to the genuine matches, and H_1 to the impostors, we can use the Neyman-Pearson statistical test [18] to optimise the classification decision in function of a threshold λ (11)

$$S = \begin{cases} 0, & \text{if } P(S|H_1) > \lambda P(S|H_0) \\ 1, & \text{if } P(S|H_0) \leq \lambda P(S|H_1) \end{cases} \tag{11}$$

Class density distributions $P(S/H_0)$ was estimated through $\hat{i}$ for positive identification (12), and $\hat{\hat{i}}$ on the negative approach (13), from class predictions ω_i

$$\hat{i} = \arg_i \max P(\omega_i|\eta) \tag{12}$$

$$\hat{\hat{i}} = \arg_i \min P(\omega_i|\eta) \tag{13}$$

Computing Bayes error rate (14) for both identification modes at varying λs, we obtain the receiver operating characteristic (ROC) curve at Fig. 10. This graphic we illustrate the performance of both identification modes by plotting the true positive rate against the false positive rate for various λ-values. A point closer to the origin (0, 0) corresponds to an higher λ-value and, consequently, a more restrictive system. We can see that relaxing the parameter λ makes true positives increase at an higher rate on negative identification than on positive identification. In the latter, true positive never gets over 0.02, which is understandable since we are using image entropy as the single feature. Nonetheless, we can see how such a poor decision environment built form a single feature, which do not provide enough information to attain positive identification,

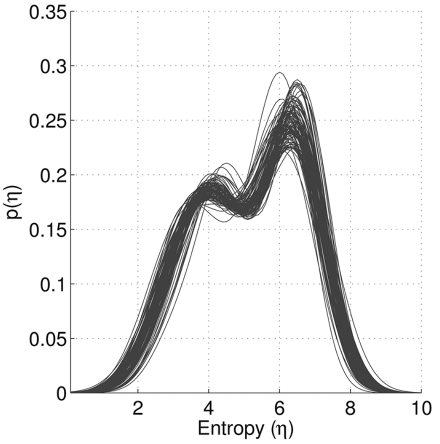


Fig. 9 *Probability density function for entropy values (η) on all subjects on the dataset*

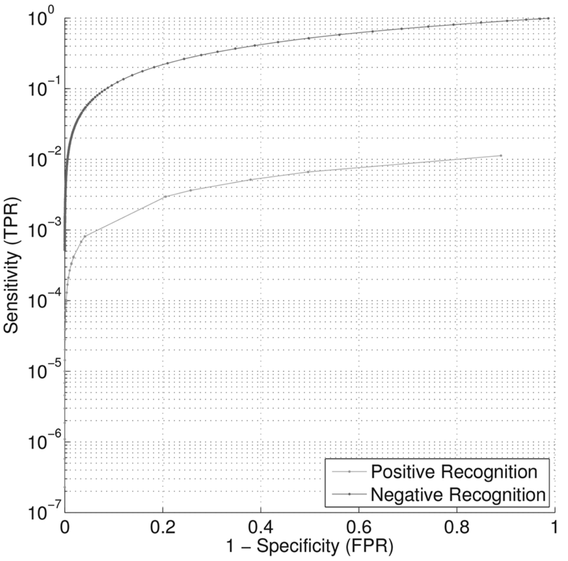


Fig. 10 *ROC curve for positive and negative identifications, using Neyman-Pearson criterium with different λ values*

still allows reliable negative identification

$$P(\text{error}|S) = \sum_{\omega_i \neq \omega_{\max}} \int_{\eta \in H_i} P(\eta|\omega_i) p(\omega_i)\, \mathrm{d}\eta \qquad (14)$$

4 Conclusions

This paper introduced the BioHDD, a new multi-session dataset of heavily images, with two singularities that turn it suitable for evaluating biometric recognition methods in extremely degraded data: (1) it contains a set of profile and frontal mugshots from 101 subjects, simulating good quality enrolment data; (2) it contains large sets of probes degraded under combinations of ten types of noise factors, resulting in images that are extremely hard to classify.

Further, we conducted an extensive on-line survey on the BioHDD data. Participants were asked to positively/negatively identify probes against the enrolled identities, along with a description of the major features used in their responses. The analysis of identification performance showed that humans have no issues cooping with inadequate illumination intensity and moderate levels of occlusions. Also, a notable ability to cope with low-resolution and compressed images was observed, suggesting that humans mostly rely on global features for identification tasks. On the other side, probes with subjects looking straight up or down and higher levels of occlusion were found to be stressful elements. That is probably the most concerning issue, as subjects trying to avoid detection ‘in the wild’ are more likely to be caring headgear or facing down, away from visible cameras.

A second level analysis was carried out on the justifications that participants gave for their responses: we concluded that high-frequency information, although not latent to the identification process, is taken into account when looking for specific attributes than can support their decisions. In both cases, shape related cues were the most accounted for, and also the more reliable. On the other side, texture information was rarely indicated as a decisive element. Holistic features, although not the more reliable ones, were also used on most justifications. From the identified features, the more reliable were the ones located on the mid-face: periocular features, the nose and the ears.

As further lines of work, authors plan to: (1) extend the acquisition setup in order to make it even more complete at mimicking ‘in the wild’ conditions (e.g. complement it with different light source angles); (2) expand the BioHDD dataset with a larger amount of participants, increasing even more the statistical significance of the dataset. Such improvements will be made available at the database website.

5 Acknowledgments

The authors would like to acknowledge the financial support provided by *FCT – Fundação para a Ciência e Tecnologia* through the research grant SFRH/BD/80182/2011, and the funding from ‘FEDER – QREN – Type 4.1 – *Formação Avançada*’, subsidised by the European Social Fund and by Portuguese funds through ‘MCTES’.

6 References

1 Jain, A.K., Pankanti, S., Prabhakar, S., Hong, L., Ross, A.: ‘Biometrics: a grand challenge’. Proc. 17th Int. Conf. on Pattern Recognition (ICPR), 2004, vol. 2, pp. 935–942

2 Ricanek, K., Savvides, M., Woodard, D.L., Dozier, G.: ‘Unconstrained biometric identification: Emerging technologies’, *Computer*, 2010, **43**, (2), pp. 56–62

3 Matey, J.R., Naroditsky, O., Hanna, K., *et al.*: ‘Iris on the move: acquisition of images for iris recognition in less constrained environments’. Proc. IEEE, 2006, vol. 94, pp. 1936–1947

4 Hu, W., Tan, T., Wang, L., Maybank, S.: ‘A survey on visual surveillance of object motion and behaviors’, *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.* 2004, **34**, (3), pp. 334–352

5 Haritaoglu, I., Harwood, D., Davis, L.S.: ‘W4: real-time surveillance of people and their activities’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (8), pp. 809–830

6 Kamgar-Parsi, B., Lawson, W., Kamgar-Parsi, B.: ‘Toward development of a face recognition system for watchlist surveillance’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (10), pp. 1925–1933

7 Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: ‘Face recognition: a literature survey’, *ACM Comput. Surv.*, 2003, **35**, (4), pp. 399–458

8 Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: ‘Face recognition by humans: nineteen results all computer vision researchers should know about’, *Proc. IEEE*, 2006, **94**, (11), pp. 1948–1962

9 Gross, R.: ‘Face databases’ (Springer Verlag, 2005)

10 Phillips, P.J., Wechsler, H., Huang, J.S., Rauss, P.J.: ‘The FERET database and evaluation procedure for face recognition algorithms’, *Image Vis. Comput.*, 1998, **16**, (5), pp. 295–306

11 Sim, T., Baker, S., Bsat, M.: ‘The CMU pose, illumination and expression database’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2003, **25**, (12), pp. 1615–1618

12 Gao, W., Cao, B., Shan, S., Zhou, D., Zhang, X., Zhao, D.: ‘The cas-peal large-scale chinese face database and baseline evaluations’. Technical Report JDL-TR-04-FR-001, ICT-ISVISION, Chinese Academy of Sciences, May 2004

13 Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: ‘Multi-pie’. FG ‘08. Eighth IEEE Int. Conf. on Automatic Face Gesture Recognition, September 2008, pp. 1–8

14 Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: ‘Labeled faces in the wild: a database for studying face recognition in unconstrained environments’. Technical Report 07-49, University of Massachusetts, Amherst, October 2007

15 Matthews, B.W.: ‘Comparison of the predicted and the observed secondary structure of t4 phage lysozyme’, 1975, **405**, pp. 442–451

16 Yager, N., Dunstone, T.: ‘Worms, chameleons, phantoms and doves: new additions to the biometric menagerie’. IEEE Workshop on Automatic Identification Advanced Technologies, 2007, pp. 1–6

17 Yager, N., Dunstone, T.: ‘The biometric menagerie’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (2), pp. 220–230

18 Neyman, J., Pearson, E.S.: ‘On the problem of the most efficient tests of statistical hypotheses’, *Philos. Trans. R. Soc. London. A, Containing Pap. Math. Phys. Charact.*, 1933, **231**, pp. 289–337