Toward Covert Iris Biometric Recognition: Experimental Results From the NICE Contests

Hugo Proença, *Member, IEEE*, and Luís A. Alexandre

***Abstract*—This paper announces and discusses the experimental results from the Noisy Iris Challenge Evaluation (NICE), an iris biometric evaluation initiative that received worldwide participation and whose main innovation is the use of heavily degraded data acquired in the visible wavelength and uncontrolled setups, with subjects moving and at widely varying distances. The NICE contest included two separate phases: 1) the NICE.I evaluated iris segmentation and noise detection techniques and 2) the NICE:II evaluated encoding and matching strategies for biometric signatures. Further, we give the performance values observed when fusing recognition methods at the score level, which was observed to outperform any isolated recognition strategy. These results provide an objective estimate of the potential of such recognition systems and should be regarded as reference values for further improvements of this technology, which—if successful—may significantly broaden the applicability of iris biometric systems to domains where the subjects cannot be expected to cooperate.**

***Index Terms*—Biometrics, forensics, iris recognition, noncooperative image acquisition, visible-light data.**

I. Introduction

IRIS recognition under controlled data acquisition protocols is a relatively mature technology that has been shown to be effective in different scenarios (e.g., airport check-in and refugee registry) and in independent technology evaluation initiatives (e.g., the ICE [18]). On the other hand, the feasibility of this technology under uncontrolled data acquisition conditions raises considerable skepticism, although the possibility of reliably recognizing human beings in unconstrained conditions is considered a grand challenge, due to the evident economic and security implications that it would have in modern societies.

This paper announces and discusses the results of the best participants in the Noisy Iris Challenge Evaluation (NICE) contest, which was the first initiative to evaluate iris recognition strategies in unconstrained data, that is, using images acquired in the visible wavelength spectrum, from widely varying distances (four to eight meters), under uncontrolled lighting conditions with moving subjects and without requiring any active participation of the subjects in the acquisition process. Such recognition systems will broaden the applicability of biometric technology to domains where the subjects cannot be expected to cooperate (e.g., surveillance and forensics), which motivates the wide interest in such systems on the part of the research community. There are presently several research programs for this purpose, and vast human and economic resources have been devoted to this task.

It is clear that data acquired in uncontrolled scenarios pose new challenges to the pattern recognition process. One of the main goals behind the NICE contest was to obtain performance measures that could be regarded as reference values in evaluating subsequent technological improvements. Moreover, compared with previous iris recognition evaluation initiatives, the NICE has two distinguishing features.

1) It uses data acquired in the visible wavelength spectrum. In order to acquire iris data from large distances, the acceptable depth-of-field values demand significantly higher f-numbers for the optical system, in direct proportion to the square of the intensity of light required for the process. Additionally, the motion factor demands very short exposure times, which again require unreasonably high levels of light. The American and European standards councils ([2] and [3]) proposed safe irradiance limits for NIR illumination of near 10 mW/cm^2, which are particularly hazardous for NIR wavelengths because the eye does not instinctively respond with its natural mechanisms (aversion, blinking and pupil contraction).
2) It was divided into two separate phases, in order to prevent errors in segmentation from biasing further processing and impeding the fair comparison between strategies. The first part of the contest (NICE.I) exclusively evaluated iris segmentation strategies and independently assessed this crucial task. Later, using data that were automatically segmented by the method that outperformed the NICE.I, the NICE:II exclusively evaluated the iris encoding and matching strategies, guaranteeing that all methods operate with the exact same data and obtaining unbiased performance measures.

The remainder of this paper is organized as follows: Section II provides an overview of the contest protocol and describes the data sets that were used. Sections III and IV provide and discuss the results obtained by the best participants in NICE.I and NICE:II. Finally, the overall conclusions are given in Section V, along with a discussion of the major trends behind this type of recognition systems.

Manuscript received August 19, 2011; revised November 09, 2011; accepted November 21, 2011. Date of publication November 29, 2011; date of current version March 08, 2012. This work was supported in part by "FCT-Fundação para a Ciência e Tecnologia" and "FEDER" in the scope of the PTDC/EIA/103945/2008 Research Project "NECOVID:Negative Covert Biometric Recognition". The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ajay Kumar.

The authors are with the Department of Computer Science, IT-Instituto de Telecomunicações, SOCIA-Soft Computing and Image Analysis Group, University of Beira Interior, 6200-Covilhã, Portugal (e-mail: hugomcp@di.ubi.pt; lfbaa@di.ubi.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2011.2177659

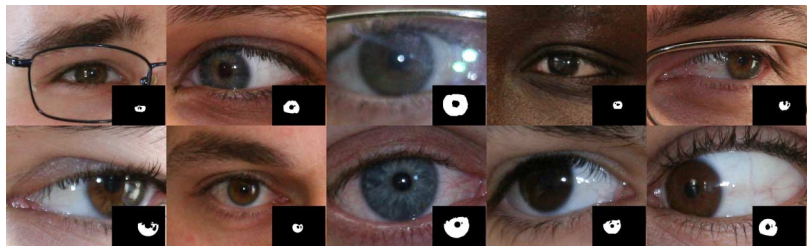


Fig. 1. Examples of degraded images used in the NICE contest, acquired from widely varying distances (three and ten meters) from moving subjects. In the bottom-right corner of each image, we plot its iris segmentation mask, which discriminates between the noise-free regions of the iris and the remaining data.

II. Contest Protocols

Both phases of the contest were free of charge and were broadly announced to research and academic communities. Each participant received a set of training data to be used in the development of their segmentation (NICE.I) or recognition (NICE:II) strategies. Participants were required to build binary executables that would run in "Microsoft Windows XP, Service Pack 2" or "Fedora Core 6" operating systems, according to a prespecified set of command line arguments. These binary executables were required to output a segmented image (NICE.I) or a text file containing a numeric dissimilarity value (NICE:II), and they had to be received by the Organizing Committee by September 30th, 2008 (NICE.I) and June 30th, 2010 (NICE:II). Finally, the submitted approaches were executed against unseen test data sets and ranked according to the observed performance. The best participants in each part of the contest were invited to describe their approaches in two special issues of the *Image and Vision Computing* (NICE.I [19]) and *Pattern Recognition Letters* (NICE:II [20]) journals. Additional information about the contest protocols can be found. [1, 2]. Also, it should be stressed that the computational complexity of the methods evaluated in both parts of the contest was not a concern and was not evaluated at any point. Actually, the comparison between execution speed might be unfair, as executables were build using different programming languages and for different operating systems. Due to this, the time-cost of each strategy is not discussed in this paper, even though significant variations among the tested strategies were observed.

A. Data Sets

The imaging framework used in the acquisition of the UBIRIS.v2 data set was installed in a lounge under both natural and artificial lighting sources. We placed several marks on the floor (between three and ten meters away from the acquisition device) and performed two distinct acquisition sessions, each lasting two weeks and separated by an interval of one week. From the first to the second session, the location and orientation of the acquisition device and artificial light sources was changed. A large majority of the volunteers were Latin Caucasian (approximately 90%), but they also included black (8%) and Asian people (2%). Approximately 60% of the volunteers participated in both imaging sessions, whereas 40% participated exclusively in one or the other. Subjects were asked to walk at a slightly slower than normal speed and to look at several lateral marks that obliged them to rotate their head and eyes, enabling the manual capture of three images per meter, between eight and four meters, giving a total of 15 images per eye and session. It should be stressed that we requested this cooperative behavior for the unique purpose of maximizing the number of usable images per subject and imaging session. A completely covert procedure could have been used, with a necessarily lower number of usable images per session. Additional details on the acquisition protocol can be found in [21]. Examples of the data used are given in Fig. 1, along with the corresponding iris segmentation mask at the bottom right corner of each figure. Also, for the purpose of reproducibility of the results given in this paper, both the training and evaluation data sets are publicly available.[3]

B. NICE.I

The first part of the contest aimed to answer the following question: "is it possible to automatically segment a small target as the iris in unconstrained data?" Ninety-seven research laboratories from 22 countries registered in the contest[4] and received a set of 500 images to be used as training data in the development of their methods. The images were 400×300 in resolution, stored in "tiff" format with 8 bits for intensity. Additionally, a set of 500 binary iris segmentation masks were manually constructed by the Organizing Committee and sent to participants to act as ground-truth data and enable automatic evaluation. Obviously, all participants agreed that such manually made ground-truth data actually correspond to the perfect iris segmentation, which in some circumstances may be a fuzzy notion.

The task assigned to participants was clear and is illustrated in Fig. 2. They had to develop a binary executable that analyzes an iris sample and outputs the corresponding binary iris segmentation mask, that is, that discriminates between the unoccluded regions of the iris (represented by white pixels) and the remaining data (black pixels). In the evaluation, disjoint test sets of 500 images were used to measure the pixel-by-pixel agreement between the segmentation masks made by each participant and the ground-truth data. The classification error rate of the jth participant was given by the average proportion of correctly classified pixels

$$E_j = \frac{1}{n\ w\ h} \sum_{i=1}^{n} \sum_{r=1}^{h} \sum_{c=1}^{w} P_i(r,c) \otimes G_i(r,c) \qquad (1)$$

[1]http://nice1.di.ubi.pt.

[2]http://nice2.di.ubi.pt.

[3]http://www.di.ubi.pt/hugomcp/doc/NICEdata.zip.

[4]http://nice1.di.ubi.pt/registered.htm.

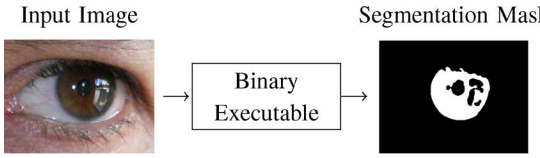


Fig. 2. Fundamental task of the NICE.I iris segmentation contest. Participations received a degraded iris image acquired in the visible wavelength (VW) and had to output the corresponding segmentation mask that discriminates between the noise-free regions of the iris and any other data.

where n is the number of images in the test set, w and h are the image width and height, respectively, $P_i(r,c)$ denotes the intensity of the pixel at line r and column c of the ith segmentation mask, $G_i(r,c)$ is the corresponding ground-truth value, and $\otimes$ is the logical exclusive-or operator. The participants were ranked according to their E_j values, and the eight best participants (listed in Table I) were invited to describe and publish their approaches in a special issue of the *Image and Vision Computing* journal [19]. The top performing method is that from the Chinese Academy of Sciences (Tan *et al.* [26]), which used a clustering-based iris localization scheme to perform a rough iris localization followed by an integro-differential constellation approach for fine pupillary and scleric border detection, which not only accelerates the traditional integro-differential operator but also enhances its global convergence. Finally, parametric models are learned to deal with eyelids and eyelashes. The entry from the Technical University of Lodz (Sankowski *et al.* [22]) localized and filled the reflections in a YIQ (luma-chrominance) color space and concluded that this approach has significant benefits for the subsequent processing. Then, it models iris boundaries by Daugman's classical integro-differential operators, followed by a parametric modeling of both eyelids. The third entry [1] presented a knowledge-based approach inspired by the expert system paradigm, which directly encodes a set of "decision rules". Authors from the Heilongjiang University (Li *et al.* [13]) performed a rough, fast eye detection and further normalized their region of interest using a c-means clustering technique. Their subsequent processing combines traditional iris segmentation techniques with RANSAC-like techniques. Authors from Dongguk University (Jeong *et al.* [11]) used an Adaboost-based technique to roughly localize the iris and to compensate for errors that result from the circular modeling of both iris borders. Later, they used color information to detect reflections and proposed a classification model to decide whether the eye is closed or not. The approach proposed by the Florida International University team (Chen *et al.* [4]) relied on the major difference in appearance between the sclera and the remaining parts of the eye to perform a coarse initial detection of the eye region. This region is used as the ROI of further processing stages, which makes their task significantly easier. They also proposed an interesting circle correction strategy to improve the segmentation results. The entry from the University of Milan (Scotti and Labbati [9]) significantly reduced the region of interest, starting with a coarse estimation of the iris and pupil centers. Later, they constrained the search for the iris boundaries exclusively within small stripes of the image, by means of iris linearization. They removed eyelashes and reflections exclusively in these stripes and finally remapped the resulting boundaries into the original domain. Finally, the approach from the Universidad Politecnica de Madrid (Luengo-Oroz *et al.* [15]) proposed a very original strategy that detects the iris center through projection techniques. The center is used to translate the region of interest into a polar coordinate system, where morphological operators are used to roughly segment the outer iris border, which is then projected back to the Cartesian space in order to suppress the eyelids and eyebrows.

TABLE I
NICE.I CLASSIFICATION

Rank	Authors	Affiliation	Error (E_j)
1	Tan *et al.* [26]	Chinese Academy of Sciences	0.0131
2	Sankowski *et al.* [22]	Technical University of Lodz	0.0162
3	P. Almeida [1]	University of Beira Interior	0.0180
4	Li *et al.* [13]	Heilongjiang University	0.0224
5	Jeong *et al.* [11]	Dongguk University	0.0282
6	Chen *et al.* [4]	Florida International University	0.0297
7	Scotti and Labbati [9]	University of Milan	0,0301
8	Luengo-Oroz *et al.* [15]	Polytechnical University of Madrid	0.0305

C. NICE:II

The second phase of the contest complemented its predecessor in terms of the traditional pattern recognition stages, evaluating different signature encoding and matching strategies. In order to guarantee that unbiased performance measures were obtained, all the participants used the exact same segmented data, which were automatically obtained according to the highest performing method in the NICE.I. Again, participation in NICE:II was free of charge and opened to any type of research and academic institution. Sixty-seven participants from 30 countries registered in the contest[5] and received a training set composed of 1000 images and the corresponding binary iris segmentation masks.

The task assigned to participants is illustrated in Fig. 3: to construct a binary executable that receives (by command-line parameters) a pair of iris samples and their iris segmentation masks and outputs a text file containing a score that corresponds to the dissimilarity between the irises. This score d should be a metric, i.e., it should meet the following conditions: 1) $d(I,I) = 0$; 2) $d(I_1,I_2) = 0 \Rightarrow I_1 = I_2$; and 3) $d(I_1,I_2) + d(I_2,I_3) \geq d(I_1,I_3)$. In the evaluation, disjoint sets of unseen 1000 images and the corresponding segmentation masks were used to rank participants. Let $\mathbb{I} = \{I_1,\ldots,I_n\}$ be a set of iris images, $\mathbb{M} = \{M_1,\ldots,M_n\}$ their binary iris segmentation masks and $id(.)$ the identity function on an image. An *one-against-all* comparison scheme yields a set of *match* $\mathbb{D}^I = \{d_1^i,\ldots,d_m^i\}$ and of *nonmatch* $\mathbb{D}^E = \{d_1^e,\ldots,d_k^e\}$ dissimilarity scores, respectively, for the cases where $id(I_i) = id(I_j)$ and $id(I_i) \neq id(I_j)$. As suggested by Daugman [6], for two-choice decisions (e.g., *match/nonmatch*) the decidability index d' measures how well

[5]http://nice2.di.ubi.pt/registered.htm.

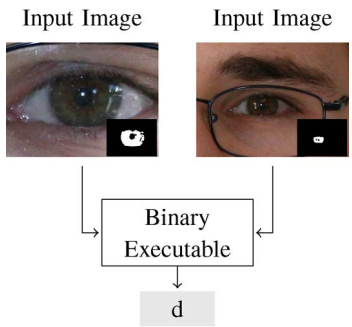


Fig. 3. Fundamental task of the NICE:II iris recognition contest. Participants must produce a binary executable that receives a pair of VW iris images and their segmentation masks and outputs a numerical value that gives the dissimilarity between the irises.

TABLE II
NICE:II CLASSIFICATION

Rank	Authors	Affiliation	Decid. (d')
1	Tan *et al.* [27] (CASIA)	Chinese Academy of Sciences	2.5748
2	Wang *et al.* [28] (NU)	Techshino Biometrics Research Center, Northeastern University	1.8213
3	Santos and Hoyle [23](UBI)	University of Beira Interior and Universidade Federal do Rio de Janeiro	1.7786
4	Shin *et al.* [24] (BERC)	Biometrics Engineering Research Center, Dongguk University	1.6398
5	Li *et al.* [14] (Peihua)	Heilongjiang University	1.4758
6	Marsico *et al.* [17](BIPLab)	University of Salerno	1.2565
7	Li and Ma [16] (HLJUCS)	Heilongjiang University	1.1892
8	Szewczyk *et al.* [25] (TUL)	Technical University of Lodz	1.0931

separated the two types of distributions are, and recognition errors correspond to their overlap area

$$d' = \frac{|\mu_E - \mu_I|}{\sqrt{\frac{1}{2}\left(\sigma_I^2 + \sigma_E^2\right)}} \tag{2}$$

where $\mu_I = (1/k)\sum_i d_i^I$ and $\mu_E = (1/m)\sum_i d_i^E$ are the means of the two distributions and $\sigma_I = (1/(k-1))\sum_i (d_i^I - \mu_I)^2$ and $\sigma_E = (1/(m-1))\sum_i (d_i^E - \mu_E)^2$ their standard deviations. The participants were ranked according to their decidability scores, and the best eight (listed in Table II) were invited to publish their approaches in a special issue of the *Pattern Recognition Letters* journal.

The best performing approach came from Tan *et al.* [27] that performed biometric recognition according to both iris and periocular data. Global color-based features and local ordinal measures were used to extract discriminating data from the iris region, later fused to periocular data extracted from texton representations. Finally, fusion is performed by the sum rule using the normalized scores generated for the different types of features. Wang *et al.* [28] used an adaptive boosting algorithm to build a strong iris classifier learned from a set of bidimensional Gabor-based set of features, each corresponding to a specific orientation and scale and operating locally. Later, given the fact that the pupillary boundary is especially difficult to segment in VW data, the authors trained two distinct classifiers: one for irises deemed to be accurately segmented and another for cases in which the pupillary boundary was not accurately segmented. Santos and Hoyle [23] fused a set of recognition techniques that can be divided in two main categories: wavelet-based textural analysis methods applied to the iris region, complemented by distribution-based (histogram of oriented gradients and local binary patterns) and scale invariant feature transforms that analyze the periocular region, which was recently suggested as an important addition for handling degraded samples, essentially because it is less vulnerable to problems resulting from deficient illumination or low-resolution acquisition. Shin *et al.* [24] started by classifying the left and right eyes by their eyelash distributions, which they used to reduce the search space. Further, they coupled two encoding and matching strategies based in color and textural analysis to obtain multiple distance scores fused by means of a weighted sum rule, which is claimed to improve the separation between *match* and *nonmatch* distributions. Li *et al.* [14] used a novel weighted co-occurrence phase histogram to represent local textural features. This method is claimed to model the distribution of both the phase angle of the image gradient and the spatial layout, which overcomes the major weakness of the traditional histogram. A matching strategy based on the Bhattacharyya distance measures the goodness of match between irises. Finally, the authors concluded that the performance is improved when a simple image registration scheme accounts for the image deformation. Marsico *et al.* [17] proposed the use of implicit equations to approximate both the pupillary and limbic iris boundaries and perform image normalization. Next, they exploited local feature extraction techniques such as linear binary patterns and discriminable textons to extract information from vertical and horizontal bands of the normalized image. Li and Ma [16] introduced an image registration method based on the Lucas-Kanade algorithm to account for iris pattern deformation. Operating on the filtered iris images, this method divides the images into small subimages and solves the registration problem for each small subimage. Later, a sequential forward selection method searches for the most distinctive filters from a family of Gabor filters, concluding that a very small number of selected features is able to obtain satisfactory performance. Finally, Szewczyk *et al.* [25] presented a semi-empirical approach based on a reverse bi-orthogonal dyadic wavelet transform, empirically selecting a compactly supported bi-orthogonal spline wavelet for which symmetry is possible with FIR filters and three vanishing moments. The authors concluded that such a method produces a short biometric signature (324 bits) that can be successfully used for recognition under such challenging conditions, improving its reliability.

III. PERFORMANCE EVALUATION: NICE.I

Fig. 4 illustrates the results obtained by the best method of the NICE.I contest, where red and green iris pixels denote Type

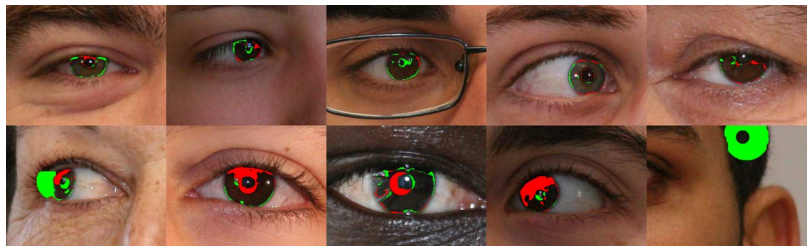


Fig. 4. Examples of the segmentation masks output by the method that got the best results in the NICE.I. Images in the upper row illustrate successful cases, whereas images in the bottom show failures. Pixels in green and red denote classification errors of Type I and II, i.e., pixels that were erroneously classified as noise-free iris pixels and pixels that were erroneously classified as not being noise-free iris pixels.

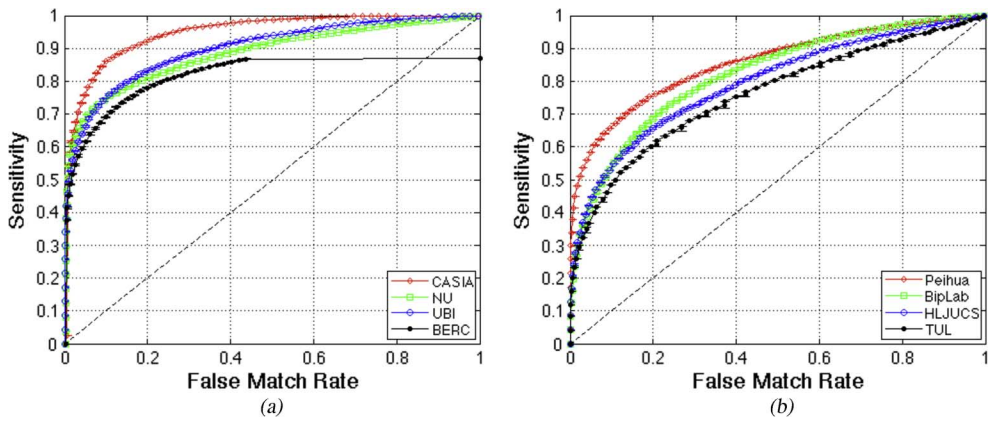


Fig. 5. Comparison between the ROC curves obtained by the best participants in the NICE:II contest. (a) ROC curves of the first to fourth participants. (b) ROC curves of the fifth to eighth participants.

I and Type II classification errors. Images in the upper row illustrate successful cases where the noise-free iris regions were accurately discriminated. Oppositely, images in the bottom row illustrate less successful cases, where evident disagreement was observed between the automatically obtained and manually constructed segmentation masks. Apart from the cases where the image did not actually contain an eye (e.g., the image at far right), we observed that most of the errors were due to difficulty in accurately segmenting the iris pupillary boundary, as confirmed by the frequent higher density of red/green pixels adjacent to the pupillary region. This occurred in a considerable number of images, which led us to conclude that—unlike in the NIR iris data—the segmentation of the pupillary boundary of VW iris samples is more difficult to perform than of the limbic boundary, which is especially evident for the heavily pigmented iris. Even so, for the vast majority of cases, a plausible segmentation of the noise-free iris regions was achieved by the best participants, which suggests that the best approaches dealt appropriately with varying acquisition distances, different amounts of light, gaze and multiple iris occluding factors (e.g., eyelids, eyelashes, glasses, contact lenses and hair).

As the outputs of the data segmentation phase are supplied as input of the subsequent processing phases (data encoding and matching), most of the performance indicators given in this paper regard the second part of the contest (NICE:II), which also integrates and indirectly evaluates the effectiveness of segmentation approaches. Also, in Section IV-D we analyze the performance of the best encoding and matching strategies with respect to the segmentation algorithm used. This way, it is possible to perceive the impact of the iris segmentation phase in the performance of the complete recognition system.

IV. Performance Evaluation: NICE:II

A. Verification Mode

In the verification mode, the performance was mainly expressed in terms of receiver operating characteristic (ROC) curves, which show the tradeoff between the two types of errors by plotting the true positive (TP) rate against the false positive (FP) rate for varying acceptance thresholds. According to the *one-against-all* comparison scheme described above, the TP rate (sensitivity) is given by the fraction of *match* dissimilarity scores that are less than or equal to the threshold t: $\mathrm{TP(t)} = (|\{d^i_\cdot \leq t\}|)/m$. Similarly, the FP rate is the fraction of *nonmatch* scores that are less than or equal to t: $\mathrm{FP(t)} = (|\{d^e_\cdot \leq t\}|)/k$, where $|.|$ denotes cardinality. In our experiments, we used $n = 1\,000$ and constructed ten different versions of the I data set, each one composed of randomly drawn images from the initial set of 14 000 images. Fig. 5 plots the error curves of the ROCs obtained for the best participants in NICE:II, where each data point is surrounded by two vertical bars that denote the best and worst values obtained at that operating point. The performance achieved by the *CASIA* algorithm is significantly better than that of the others at any operating point, whereas the curves of the other approaches often intersect, suggesting that each algorithm outperforms the others in a specific operating range.

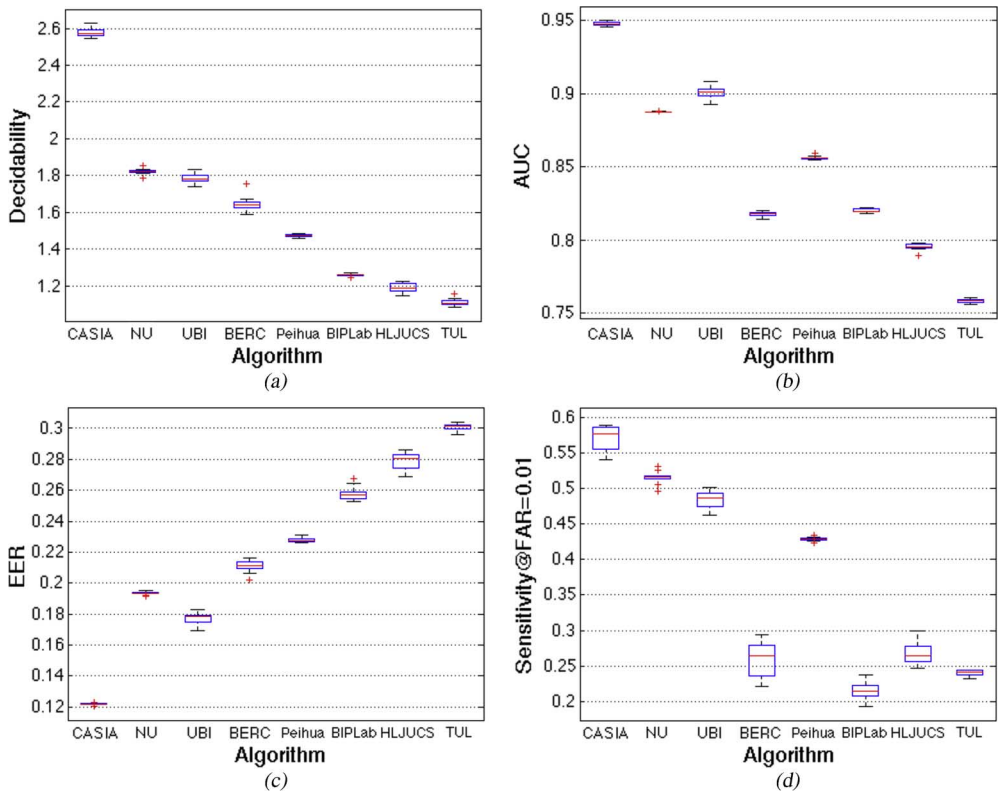


Fig. 6. Summary of the performance measures obtained by the best participants in the NICE:II contest. Each column in the graphs reports the performance value of a participant, where the median value is represented by the horizontal line through the middle of each box. Top and bottom of the boxes denote the first and third quartile of the observations. Outliers appear as dot data points. (a) Boxplot of the algorithms' decidability; (b) boxplot of the algorithms' area under curve (AUC); (c) boxplot of the algorithms' equal error rate (EER); (d) boxplot of the algorithms' Sensitivity@FAR $= 0.01$.

In Fig. 6, we compare the most common performance measures obtained for the best algorithms in NICE:II. The results are expressed in terms of boxplots, showing the median of the observed performance range (horizontal solid line) and the first and third quartile values of the observations (top and bottom of the box marks). The upper and lower whiskers are denoted by the horizontal lines outside each box, and the outliers are denoted by dot points. Fig. 6(a) shows the contest rankings, displaying the observed decidability values. Fig. 6(b) and (c) provides complementary overall performance measures (AUC and EER), that are often correlated, which was confirmed in our observations. Fig. 6(d) summarizes the sensitivity values observed when operating at a FAR of approximately 0.01. Two algorithms (CASIA and NU) met or exceeded the 0.5 value, meaning that they correctly recognized more than half of the subjects. Overall, the different ranges of the participants' boxplots highlight the small variance of two of them (NU and PeihuaLi), according to the different data sets. Although they are ranked third and fifth, this suggests that a better tradeoff bias/variance could have been obtained, and thus, a significant margin for improvements exist. The approach of CASIA has consistently outperformed all the other performance measures, but the relative ranks for each performance measure were not stable among the other algorithms. These results give clear evidence that these recognition strategies were able to extract and match discriminating information even from such degraded data, which is a positive result and encourages further improvements.

B. *Identification Mode*

Using a closed universe model, we tested the effectiveness of each algorithm when trying to answer the following question: "is the correct answer among the best k matches?" This type of performance measure is usually expressed by means of rank and cumulative rank histograms, where ranks appear in the horizontal axis and probabilities in the vertical one. Let $\mathbb{T} = \{T_1, \ldots, T_t\}$ be the set of gallery images such that $i \neq j \Rightarrow id(T_i) \neq id(T_j)$ and $\mathbb{S} = \{S_1, \ldots, S_s\}$ the set of samples that are to be compared against $\mathbb{T}$. Each S_i constitutes a query that is matched against all elements of $\mathbb{T}$, yielding a set of $D = \{d_{i1}, \ldots, d_{it}\}$ dissimilarity scores, where d_{ij} denotes the dissimilarity between the ith sample and the jth template. Let $D' = \{d'_{i1}, \ldots, d'_{it}\}$ be the ordered version of D, such that $d'_{i1} \leq d'_{i2} \leq \ldots \leq d'_{it}$. S_i is said to have rank k if the score d_{ij} is in the kth position of D' and $id(S_i) = id(T_j)$. The probability of having rank k P(rank-k) is estimated by the ratio between the number of sample queries with rank k and the total number of queries. Accordingly, the probability of cumulative rank k can be calculated as $\sum_{i=1}^{k}$ P(rank-i).

Fig. 7 gives the probability distribution of the cumulative rank values obtained for $t = 100$, representing the identification performance obtained. Interestingly, the relative performance of the algorithms has some differences compared to the contest classification: the identification performance of UBI (third) and BERC (fourth) was consistently better than that of NU (second).

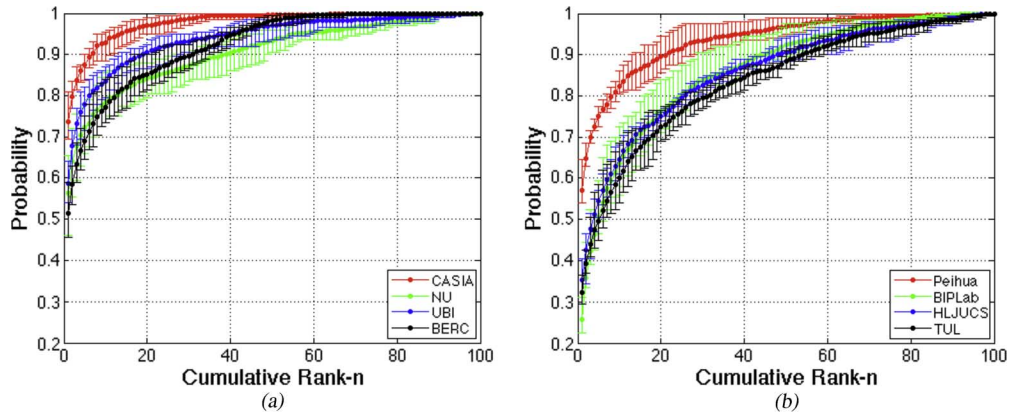


Fig. 7. Average cumulative rank n curves obtained by the best participants in the NICE:II. Bottom and top horizontal lines around the data series denote the worst and best values obtained. (a) Average cumulative rank n curves of the first to fourth participants; (b) average cumulative rank n curves of the fifth to eighth participants.

TABLE III
PEARSON'S SAMPLE CORRELATION COEFFICIENTS BETWEEN NICE:II BEST PARTICIPANTS. PARTICIPANTS ARE DENOTED BY CORRESPONDING RANK

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1.00	–	–	–	–	–	–	–
2	0.56	1.00	–	–	–	–	–	–
3	0.69	0.58	1.00	–	–	–	–	–
4	0.65	0.45	0.52	1.00	–	–	–	–
5	0.65	0.56	0.68	0.51	1.00	–	–	–
6	0.56	0.41	0.57	0.39	0.61	1.00	–	–
7	0.55	0.46	0.65	0.44	0.64	0.41	1.00	–
8	0.48	0.43	0.58	0.40	0.53	0.39	0.52	1.00

We also observed that this type of performance measure suffers significant variations with respect to different test data sets, which is easily observed in the variance intervals around each data point. Here, the maximal variance was observed when obtaining the cumulative rank values of approximately $t/4$. Regarding the rank 1 values, the best performance was observed for the CASIA algorithm, which attained correct identifications for approximately 74% of the number of queries. On the other hand, the lowest values were observed for BIPLAB, which returned the correct identification in rank-1 for approximately 25% of the queries. Finally, when considering the cumulative rank value required to return the correct identity in all queries, large heterogeneous values were observed: the best value was rank 37 for CASIA, followed by rank 57 for BERC. At the other extreme, BIPLAB, HLJUCS and TUL obtained values close to 100, which means that the match dissimilarity score in these situations was among the highest.

C. Correlation and Fusion of Results

The statistical correlation between the outputs given by the best participants in NICE:II was analyzed to address the question of whether performance can be improved by fusing several of them together. It was assumed that any eventual dependence between scores would be at most linear, which justifies the use of the Pearson's correlation coefficient to analyze the strength of these dependences. Table III gives the correlation scores, given by $r(X,Y) = (1/(n-1))\sum_i((X_i - \bar{X})/\sigma_X)((Y_i - \bar{Y})/\sigma_Y)$, where X_i and Y_i denote the system outputs, $\bar{X}, \bar{Y}$ are the sample means and σ_Y, σ_Y the standard deviations. The highest correlation value was observed for the {CASIA, UBI} algorithms (almost 0.7), followed by {UBI, Peihua}, {CASIA, BERC} and {CASIA, Peihua}, all of them above 0.6. On the other hand, BIPLab appeared to be the least correlated algorithm, achieving the lowest scores in {BIPLab, BERC} and {BIPLab, TUL} (around 0.39).

Using the theoretical framework developed by Kittler *et al.* [12], all the combinations of biometric experts were tested, according to the usual fusion rules: product $(*)$, sum $(+)$, min (m) and max (M). Without any assumption on the prior probabilities, the posterior probability that a pattern $\overrightarrow{x_i}$ belongs to class w_j was obtained by

$$P(w_j|\overrightarrow{x_i}) = \frac{P(\overrightarrow{x_i}|w_j)}{\sum_s P(\overrightarrow{x_i}|w_s)}. \quad (3)$$

An input pattern is assigned to class w_c if $w_c = \arg_j \max \phi P(w_j|\overrightarrow{x_i})$, where ϕ denotes the combination rule. Table IV lists the best classification ensembles obtained according to the number of fused experts (column #Fused) and the best combination rule observed (column CR). The methods are denoted by their rank in the contest. Also, Fig. 8 reports the EER, AUC and sensitivity performance values of the best classification ensembles, when compared to the best individual method. A slight but consistent improvement in performance due to fusion was observed, and the maximal performance attained when the best four experts were fused according to the product rule.

D. Sensitiveness to Segmentation

Data segmentation is a key phase in most pattern recognition systems, as errors easily bias the feature encoding and matching strategies. Additionally, as it is one of the earliest phases, the segmentation module should more directly handle the dynamics of the acquired data, which is particularly important for the type of biometric recognition that is the focus of this paper. Due to the above points, we found it meaningful to assess how much the recognition performance may be affected by errors in segmentation in order to obtain an idea of how further improvements in segmentation would improve the overall system performance. Thus, we compared the performance obtained by the best participants in the NICE:II when using *not so well* segmented data, i.e., using the segmentation results of the second,

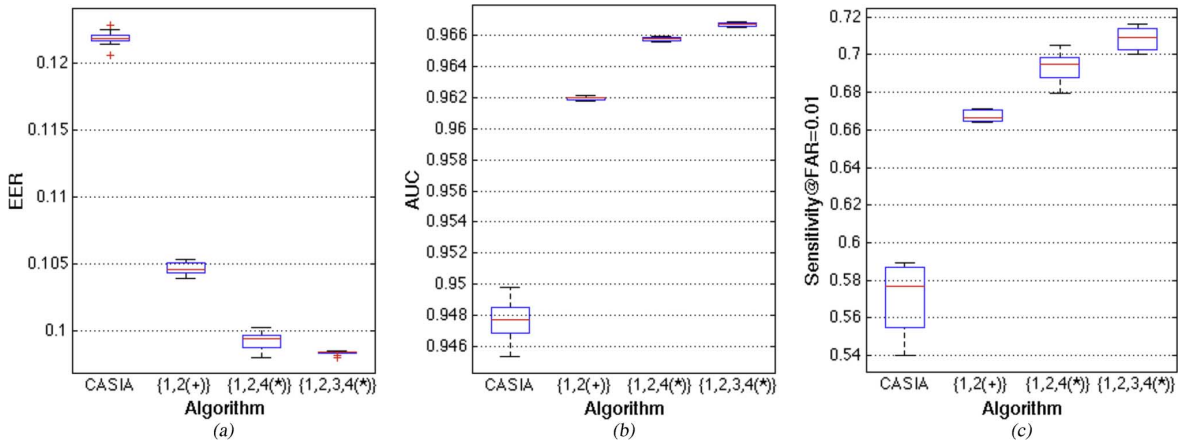


Fig. 8. Comparison between the performance obtained by the best method of NICE:II and by the best classification ensembles of two, three and four classifiers. (a) EER; (b) AUC; (c) Sensitivity@FAR $= 0.01$.

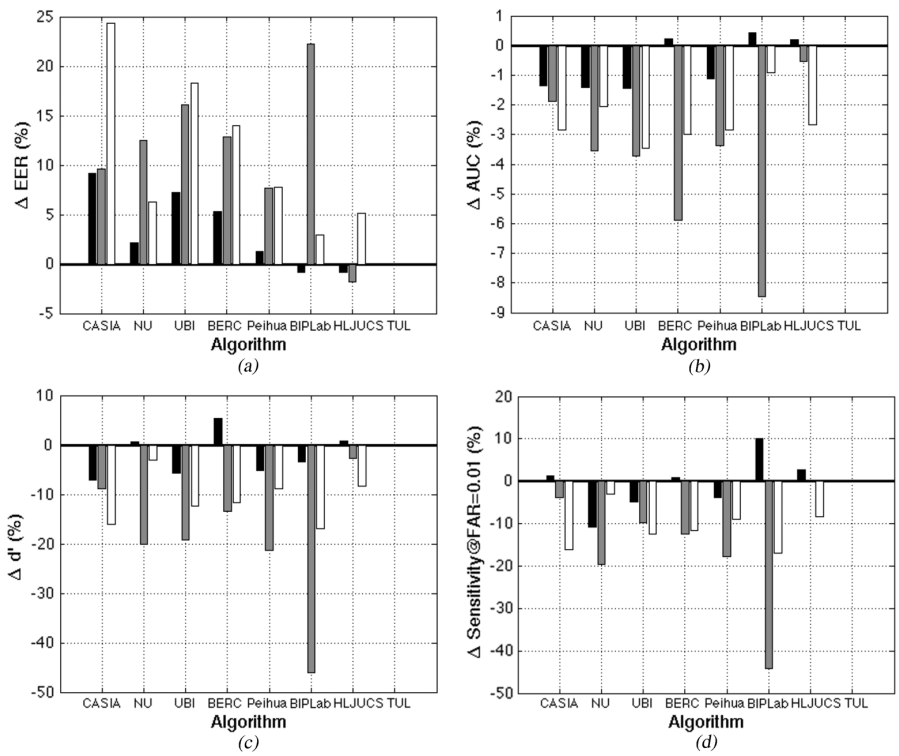


Fig. 9. Percentile variations in performance of the best participants of NICE:II when the best segmentation method of NICE.I was not used, but rather data segmented by the second (black bars), third (gray bars) and fourth (white bars) best segmentation algorithms. (a) EER; (b) AUC; (c) decidability; (d) Sensitivity@FAR $= 0.01$.

TABLE IV
BEST RESULTS OBTAINED BY CLASSIFICATION ENSEMBLES, ACCORDING TO NUMBER OF FUSED METHODS. METHODS ARE DENOTED BY CORRESPONDING RANK

# Fused	Methods	AUC (ΔAUC)	CR
2	{1,2}	0.9618 (+0.0150)	+
3	{1,2,4}	0.9658 (+0.0190)	*
4	{1,2,3,4}	0.9669 (+0.0201)	*
5	{1,2,3,4,6}	0.9659 (+0.0191)	*
6	{1,2,3,4,6,8}	0.9641 (+0.0173)	*
7	{1,2,3,4,5,6,8}	0.9617 (+0.0149)	*
8	{1,2,3,4,5,6,7,8}	0.9592 (+0.0124)	*

third and fourth participants of the NICE.I phase. Fig. 9 illustrates these results: bars of different intensity denote the percentile variation in results when using the second (black bars), third (gray bars) and fourth (white bars) best NICE.I segmentation methods. The horizontal solid lines denoting the results obtained when using the best segmentation method, for comparison. Not surprisingly, an evident deterioration in performance was observed, which usually surpasses the differences in segmentation performance, i.e., the segmentation errors were observed to propagate in increasing amplitudes for further processing stages. However, it is interesting to note that the BIPLab approach consistently achieved better performance using the data segmented by the second algorithm of NICE.I than using the best segmentation method. Finally, the performance of TUL algorithm was independent of the input segmentation masks, which suggests that this algorithm does not use such masks in data encoding and matching processes.

E. Sensitivity to Number of Gallery Images

It is known that rank and cumulative rank scores suffer significant variations according to the database size, i.e., the number of enrolled gallery identities. This is of the most interest to predict the performance of biometric systems in realistic deployment scenarios, namely, for large scale identification purposes. Hence, we started by estimating the probability that a query on a sample S_i returns a cumulative rank k value, as a function of the database size. Then, we plotted the values observed for the best classification ensemble of NICE:II against this theoretical performance line, aiming to validate our observations and to infer the performance for large scale identification scenarios. We now assume that X and Y are independent random variables that represent the dissimilarity scores generated by a biometric classifier for *match* and *nonmatch* comparisons. Even though iris match distributions are usually asymmetric and that nonmatch distributions are often skewed with longer left tails, for evaluation purposes an oversimplification was made and both distributions assumed as normal, which is more plausible in less favorable environment conditions [7]. Let $X \sim \mathcal{N}(\mu_x, \sigma_x)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y)$ denote such *match* and *nonmatch* distributions. According to the elementary theory of rank tests [10], the kth-order statistic $Y_{(k)}$ of a statistical sample $Y_1, \ldots, Y_n$ is equal to its kth smallest value. Let $Y_{(1)}, \ldots, Y_{(n)}$ be the order statistics of a set of independent observations $Y_1, \ldots, Y_n$, that is, $Y_{(1)} < Y_{(2)} < \ldots < Y_{(n)}$. The distribution function of $Y_{(k)}$ is equal to

$$\begin{aligned} F_{(k)}(y) &= P\left(Y_{(k)} \leq y\right) \\ &= \sum_{i=k}^{n} \binom{n}{i} \left[F(y)\right]^i \left[1 - F(y)\right]^{n-i}. \end{aligned} \tag{4}$$

In practical terms, (4) can be very hard to calculate for very large n values and $i \approx (n/2)$. Thus, using the concepts of the complement of an event and of the probability of nonoccurrence, (4) is equivalent to

$$F_{(k)}(y) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} \left[F(y)\right]^i \left[1 - F(y)\right]^{n-i}. \tag{5}$$

Using (5) and (6), it is possible to obtain the density of the kth smallest value of the *nonmatch* comparisons, for a given population of size n. Considering that $f(y)$ is the density of $F(y)$, the density of $Y_{(k)}$ is given by

$$\begin{aligned} f_{(k)}(y) &= n\binom{n-1}{k-1} \left[F(y)\right]^{k-1} \left[1-F(y)\right]^{n-k} f(y) \\ &= \frac{n!}{(k-1)!(n-k)!} \left[F(y)\right]^{k-1} \left[1-F(y)\right]^{n-k} f(y). \end{aligned} \tag{6}$$

The probability that a *match* distance observation X returns a cumulative rank k is given by

$$\begin{aligned} P\left(\mathrm{rank}(X) \leq \mathrm{k}\right) &= P(X \leq Y_k) \\ &= \int_{-\infty}^{\infty} f_{(k)}(y) \int_{-\infty}^{y} f(x)\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{7}$$

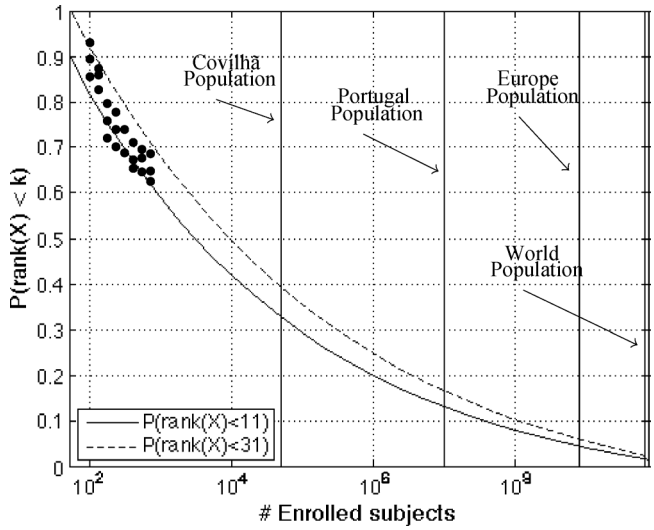


Fig. 10. Estimated probability that an identification query returns the true identity out of the first 10 (continuous line) and 30 positions (dashed line), as a function of the number of enrolled identities, i.e., $\mathrm{P(rank(X)} \leq 10)$ and $\mathrm{P(rank(X)} \leq 30)$. Circular data points are the cumulative rank 10 values observed empirically, according to the different databases sizes.

Accordingly, the probability that the *match* observation has exact rank k is given by

$$P\left(\mathrm{rank(X)} = \mathrm{k}\right) = P\left(X \leq Y_{(k)}\right) - P\left(X \leq Y_{(k-1)}\right). \tag{8}$$

The continuous line of Fig. 10 estimates the probability of a cumulative rank 10 ($P(\mathrm{rank(X)} \leq 10)$) as a function of the gallery size n, and the dashed line is analogous for a cumulative rank 30. We used $X \sim \mathcal{N}(0.32, 0.1)$ and $Y \sim \mathcal{N}(0.5, 0.06)$ values obtained by Gaussian curve fitting to the best classification ensemble composed by the fusion $(*)$ of the {CASIA, NU, UBI, BERC} algorithms (Table IV). The circular data points are the cumulative rank 10 values observed empirically, according to the different databases sizes we were able to test. For contextualization purposes, the four vertical solid lines illustrate identification scenarios at different scales: global (World population), continental (Europe), country (Portugal) and small city (Covilhã, a Portuguese city with approximately 50 000 inhabitants). From this analysis, it appears that the observed values adequately fit the probability line, which is a good indicator of the potential performance of this type of recognition system on larger scales. It should be concluded that this type of recognition technology is not yet sufficiently mature to be deployed in large scale identification scenarios, and further advances in the technology are needed to meet the full range of operational requirements at those operating scales. Currently, approximately 2% of the queries would return the correct identity in top-10, if the entire world population is enrolled in the system. This value rises to approximately 4% and 12% if the universe is reduced to the continental or national population.

F. Fusion of Multiple Recognition Systems

As stated above, the type of biometric recognition system discussed in this paper aims to operate covertly in uncontrolled data acquisition environments, meaning that absolutely no human

effort is demanded of the subjects during the recognition processes. This is a key feature that raises the possibility of using multiple recognition systems regularly spaced across an airport terminal hallway or a city street. This section discusses this possibility and how these systems could interact to improve their performance.

It is known that not all subjects perform consistently in terms of false matches and nonmatches of a biometric system. Based on their intrinsic features, some are difficult to match (*goats*), while others are particularly vulnerable to impersonation (*lambs*) and consistently increase the probability of false matches [29]. We oversimplify the problem and regard all subjects of a population $\mathbb{P} = \{s_1, \ldots, s_n\}$ as *sheep*, i.e., subjects that tend to follow the system averages: they match relatively well against themselves and poorly against others. Let us consider k iris recognition systems with roughly similar performance, with a sensitivity of α at a false match rate of β. Here, we introduce the concept of *exogenous independence*, hypothesizing that purposely changing the lighting conditions in the environment (by using different levels of light or types of illuminants) and the acquisition protocols (poses, distances) should potentiate the independence between the system outputs. Assuming that the independence of each system provides an upper bound on the performance that would be attainable by the fusion of multiple systems, the binomial distribution can be used to obtain the probability that a subject s_i is screened by k recognition systems and correctly recognized by k' of these, $1 \leq k' \leq k$:

$$P(R_{k'}) = \frac{k!}{k'!(k-k')!}\alpha^{k'}(1-\alpha)^{k-k'}. \tag{9}$$

For different values of k', the probability that a reported match is false is given by $\beta^{k'}$, assuming that false matches in each of the k recognition systems are independent events. Accordingly, a *match* will be reported *iff* a minimum of k' recognition systems output a match

$$\begin{aligned} P(R_{\geq k'}) &= \sum_{j=k'}^{k} P(R_j) \\ &= \sum_{j=k'}^{k} \frac{k!}{j!(k-j)!}\alpha^{j}(1-\alpha)^{k-j} \end{aligned} \tag{10}$$

provided that all events are mutually exclusive. Considering the median sensitivity value plotted in Fig. 8(c) for the best ensemble composed of algorithms {CASIA, NU, UBI, BERC} ($\alpha = 0.7067, \beta \approx 0.01$), Fig. 11 relates the expected sensitivity of such a *multipoint* biometric system to the number of recognition systems used, considering different false match rates. From its analysis, one can conclude that approximately five independent recognition systems would be enough to attain almost full sensitivity at a false acceptance rate β of 0.01. This value substantially increases when a lower number of false alarms is convenient (large scale applications), requiring between 13 and 23 independent recognition systems to operate, respectively, at FAR $1e^{-4}$ and $1e^{-6}$.

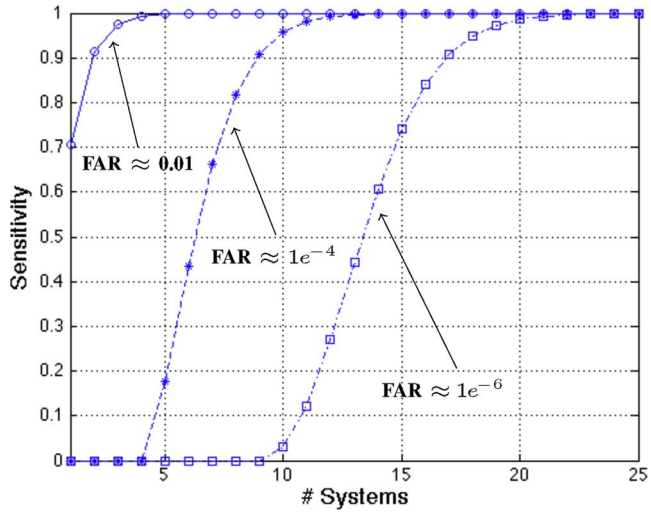


Fig. 11. Expected sensitivity of a set of iris recognition systems placed consecutively and operating covertly under intentionally varying lighting conditions, with different required values for the false acceptance rates.

V. Conclusion and Trends

We announced and discussed the experimental results from the NICE initiative, a recent iris recognition technology evaluation initiative that operates with severely degraded visible wavelength data acquired in uncontrolled protocols and scenarios. The NICE was composed of two separate phases: 1) noise-free iris segmentation (NICE.I) and 2) signatures encoding and matching (NICE:II). Its main goal was to supply performance values that can be regarded as reference values for further technology improvements. Experiments lead us to conclude that the extremely ambitious recognition systems discussed in the paper are still in an early stage of development, and that significant improvements are required, specifically to allow their application in large-scale identification scenarios. However, the economic and security implications that they would have in modern societies justifies the growing interest of the research community and presages future achievements.

Apart from improvements in each of the traditional pattern recognition stages, additional research effort should focus on addressing the most convenient ways to augment the independence between several instances of these recognition systems, taking advantage of the fact that no human effort is involved in the data acquisition process, and thus, there is no apparent evidence against the possibility of multiple recognition systems operating complementarily.

References

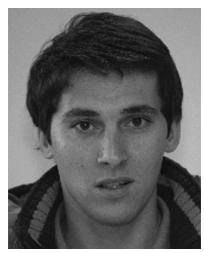
[1] P. Almeida, "A knowledge-based approach to the iris segmentation problem," *Image Vision Computing*, vol. 28, no. 2, pp. 238–245, 2010.

[2] *American National Standard for the Safe Use of Lasers and LEDs Used in Optical Fiber Transmission Systems*, ANSI Z136.2, Amer. Nat. Standards Inst., 1988.

[3] Commission International de l'Eclarirage, Photobiological Safety Standards for Safety Standards for Lamps 1999, Rep. TC 6-38; CIE 134-3-99.

[4] Y. Chen, M. Adjouadi, C. Han, J. Wang, A. Barreto, N. Rishe, and J. Andrian, "A highly accurate and computationally efficient approach for unconstrained iris segmentation," *Image Vision Computing*, vol. 28, no. 2, pp. 261–269, 2010.

[5] J. G. Daugman, "Phenotypic versus genotypic approaches to face recognition," in *Face Recognition: From Theory to Applications*. Heidelberg, Germany: Springer-Verlag, 1998, pp. 108–123.
[6] J. G. Daugman, Biometric decision landscapes Tech. Rep. TR482, 2000 [Online]. Available: http://www.cl.cam.ac.uk/users/jgd1000/biomdecis.ps
[7] J. G. Daugman, "How iris recognition works," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 21–30, Jan. 2007.
[8] J. G. Daugman, "New methods in iris recognition," *IEEE Trans. Syst., Man, Cybern.—Part B: Cybernetics*, vol. 37, no. 5, pp. 1167–1175, May 2007.
[9] R. Labati and F. Scotti, "Noisy iris segmentation with boundary regularization and reflections removal," *Image Vision Computing*, vol. 28, no. 2, pp. 270–277, 2010.
[10] J. Hajek and Z. Sidak, *Theory of Rank Tests*, 1st ed. Boston, MA: Academic, 2000.
[11] D. Jeong, J. Hwang, B. Kang, K. Park, C. Won, D. Park, and J. Kim, "A new iris segmentation method for non-ideal iris images," *Image Vision Computing*, vol. 28, no. 2, pp. 254–260, 2010.
[12] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combing classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 3, pp. 226–239, Jun. 1998.
[13] P. Li, X. Liu, L. Xiao, and Q. Song, "Robust and accurate iris segmentation in very noisy iris images," *Image Vision Computing*, vol. 28, no. 2, pp. 246–253, 2010.
[14] P. Li, X. Liu, and N. Zhao, "Weighted co-occurrence phase histogram for iris recognition," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[15] M. Luengo-Oroz, E. Faure, and J. Angulo, "Robust iris segmentation on uncalibrated noisy images using mathematical morphology," *Image Vision Computing*, vol. 28, no. 2, pp. 278–284, 2010.
[16] P. Li and H. Ma, "Iris recognition in non-ideal imaging conditions," *Pattern Recognition Lett., Special Issue on the Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[17] M. Marsico, M. Nappi, and D. Riccio, "Noisy iris recognition integrated scheme," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[18] P. J. Phillips, T. Scruggs, A. O'Toole, P. J. Flynn, K. W. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 5, pp. 831–846, Oct. 2010.
[19] H. Proença and L. A. Alexandre, Eds., "Special issue on the segmentation of visible wavelength iris images acquired at-a-distance and on-the-move," *Image Vision Computing*, vol. 28, no. 2, Feb. 2010.
[20] H. Proença and L. A. Alexandre, Eds., "Special issue on signatures encoding and matching of segmented noisy iris images," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[21] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, "The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 8, pp. 1502–1516, 2010.
[22] W. Sankowski, K. Grabowski, M. Napieralska, M. Zubert, and A. Napieralski, "Reliable algorithm for iris segmentation in eye image," *Image Vision Computing*, vol. 28, no. 2, pp. 231–237, 2010.
[23] G. Santos and E. Hoyle, "A fusion approach to unconstrained iris recognition," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[24] K. Shin, G. Nam, D. Jeong, D. Cho, B. Kang, K. Park, and J. Kim, "New iris recognition method for noisy iris images," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[25] R. Szewczyk, K. Grabowski, M. Napieralska, W. Sankowski, M. Zubert, and A. Napieralski, "Reliable iris recognition algorithm based on reverse biorthogonal wavelet transform," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[26] T. Tan, Z. He, and Z. Sun, "Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition," *Image Vision Computing*, vol. 28, no. 2, pp. 223–230, 2010.
[27] T. Tan, X. Zhang, Z. Sun, and H. Zhang, "Noisy iris image matching by using multiple cues," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[28] Q. Wang, X. Zhang, M. Li, X. Dong, Q. Zhou, and Y. Yin, "Adaboost and multi-orientation 2D Gabor-based accurate noisy iris recognition," *Pattern Recognition Lett., Special Issue Recognition of Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance*, to appear.
[29] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 2, pp. 220–230, Apr. 2010.



Hugo Proença (M'09) received the B.Sc. degree in mathematics/informatics from the University of Beira Interior, Covilhã, Portugal, in 2001, and the M.Sc. degree from the Faculty of Engineering, University of Oporto, in 2004, and the Ph.D. degree in computer science and engineering from the University of Beira Interior, in 2007.

His research interests are mainly focused in the artificial intelligence, pattern recognition and computer vision domains of knowledge, with emphasis to the biometrics area, namely the study of iris recognition systems less constrained to subjects. Currently, he serves as Assistant Professor in the Department of Computer Science, University of Beira Interior.

Dr. Proença is an author/co-author of over 40 publications, either in ISI-indexed international journals or conferences. He is member of the Editorial Board of the *International Journal of Biometrics* and served as Guest Editor of special issues of the *Pattern Recognition Letters* and *Image and Vision Computing* journals.

Luís A. Alexandre received the B.Sc. degree in physics/applied mathematics, M.Sc. degree in industrial informatics, and Ph.D. degree in electrical engineering and computers, all from the University of Porto, Portugal, in 1994, 1997, and 2002, respectively.

His research interests are pattern recognition, neural networks, image processing and biometrics. He is the author of more than 60 research papers in journals and conferences and is the Leader of a research lab at the University of Beira Interior, Portugal, where he is currently an Associate Professor.

Dr. Alexandre is a member of the Portuguese Association for Pattern Recognition, the International Neural Network Society, and has served as member of the Executive Committee (2008 to 2010) of the European Neural Network Society.