DOI: https://doi.org/10.1016/j.patcog.2017.08.028

Insights into the results of MICHE I - Mobile Iris CHallenge Evaluation

Maria De Marsico[a], Michele Nappi[b], Fabio Narducci[b,*], Hugo Proença[c]

[a] Department of Computer Science, Sapienza University of Rome, Rome, Italy
[b] Department of Computer Science, University of Salerno, Salerno, Italy
[c] IT: Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal

Keywords:
Mobile Iris Recognition
Evaluation
Biometric algorithm fusion

A B S T R A C T

Mobile biometrics technologies are nowadays the new frontier for secure use of data and services, and are considered particularly important due to the massive use of handheld devices in the entire world. Among the biometric traits with potential to be used in mobile settings, the iris/ocular region is a natural candidate, even considering that further advances in the technology are required to meet the operational requirements of such ambitious environments. Aiming at promoting these advances, we organized the Mobile Iris Challenge Evaluation (MICHE)-I contest. This paper presents a comparison of the performance of the participant methods by various Figures of Merit (FoMs). A particular attention is devoted to the identification of the image covariates that are likely to cause a decrease in the performance levels of the compared algorithms. Among these factors, interoperability among different devices plays an important role. The methods (or parts of them) implemented by the analyzed approaches are classified into seg- mentation (S), which was the main target of MICHE-I, and recognition (R). The paper reports both the results observed for either S or R, and also for different recombinations (S+R) of such methods. Last but not least, we also present the results obtained by multi-classifier strategies.

1. Introduction

Typically, benchmark datasets follow the progress of the research they help to assess. When a new research line is started, the first attempts investigate and evaluate possible solutions addressing the basic/simplest formulation of problems. Once acceptable solutions are achieved in each round, new harder issues are typically tackled. Extended and more challenging datasets are collected when the need arises to evaluate new emerging potential solutions. The iris recognition domain follows this general trend. The first version of CASIA iris dataset (CASIA-IrisV1)[1] was collected by avoiding most possible distortions that possibly hinder recognition of the iris in an image. This was not only obtained by an ideal acquisitions of the subjects [16]), but also by avoiding/bypassing segmentation difficulties, through the substitution of the pupil with a black circle. These conditions make it poorly usable at present, and tests carried out using such benchmark are scarcely significant. Notwithstanding this, the great value of this dataset is that it was among the first publicly available ones for iris, allowing a fair comparison of the first research results on the problem. Later, increasingly difficult datasets have been offered to the research community to spur and validate research on harder problems. The Mobile Iris Challenge Evaluation (MICHE)-I [7] is among the most recent competitions in the iris recognition domain, supported by the official dataset MICHE-I [8]. The aim of the competition has been to evaluate state-of-the-art algorithms to segment and encode/match iris data, captured by mobile devices in uncontrolled settings. As these devices have become massively used around the entire world, their potential for biometric recognition applications has been considered one of the major challenges for the research community. Major problems to address are due to the heterogeneity of the environments where these devices work, the different features of sensors, and the possible lack of technical experience/awareness of users, that often produce samples of unpredictable quality. Often mentioned with the term *in-the-wild*, the research in unconstrained biometric scenarios is gaining increasing attention as discusses in [25,26] for long-range iris recognition or in [28] and [29] where the typical challenges of segmentation in noisy acquisitions are addressed.

The classical solutions for iris recognition have been devised to work on data acquired in near-infrared (NIR) wavelengths. This reduces most data noise, by reducing the potential negative effect of reflections due to the cornea. Even though these algorithms are remarkably effective in noise-free data, their performance is seriously affected by the image variation factors typical in images

* Principal corresponding author.
E-mail addresses: demarsico@di.uniroma1.it (M. De Marsico), mnappi@unisa.it (M. Nappi), fnarducci@unisa.it (F. Narducci), hugomcp@di.ubi.pt (H. Proença).

[1] All CASIA datasets are available at http://biometrics.idealtest.org/

acquired in visible wavelength (VW), and in particular in mobile settings. There is an obvious need of developing new recognition solutions, particularly suitable to handle data acquired from handheld devices. Notwithstanding the advances of technology and the growing availability of computing and communication resources, the ability to transfer the overall biometric processing on a mobile device still calls for faster as well as lighter procedures, and for a smarter storage strategy. Therefore, present techniques to carry out detection, segmentation and coding, as well as matching steps, must be adapted to the mobile setting. The Mobile Iris Challenge Evaluation (MICHE)-I has intended to be an arena to compare state-of-the-art approaches to the different mobile iris processing steps. In order to provide a common ground for the comparison of proposed methods, the participants could exploit a new iris biometric dataset, namely MICHE-I,[2] captured under uncontrolled settings using mobile devices. The next sections will first present the dataset and sketch the main features of the compared methods. Afterwards some relevant aspects of the achieved results will be discussed, with a special focus on the image features/distortions that can mostly positively/negatively affect recognition performance, and on interoperability issues raising from the use of different devices i n enrollment vs. testing operations. A noteworthy aspect of the analysis carried out, has been to decompose methods proposing both a segmentation and a recognition technique, and to assemble/reassemble the obtained modules in order to investigate the best combinations. Last but not least, the paper also presents the results of possible multi-classifier strategies, to complement the strengths of different approaches. The tested fusion rules include both a very simple combination of results at score level (Simple Sum) and a more advanced technique. The latter entails to assign a different weight to the contributions by the different methods exploited. The Matcher Weighting Fusion exploits Equal Error Rate (EER) achieved by the recognition methods in a pre-testing step, and assigns a higher weight to those methods that achieve a lower EER. The weights are therefore inversely proportional to the errors of the methods considered.

The different iris/non iris segmentation strategies, have been evaluated by the classical performance measures for binary classification: Accuracy, Precision, Sensitivity, Specificity, Pratt, F1_Score, Rand Index, Global Consistency Error, E1_score, Pearson Correlation Coefficient. Final recognition, when included in the participant methods, was carried out in verification mode (1:1 matching). Each probe was compared with all the templates of a same individual in the gallery, either with the same identity of the probe (genuine attempt) or with a different identity (impostor attempt), and the best result was used to determine the system response. The performance measures used for this step have been Decidability index, Area Under Curve (AUC) and Equal Error Rate (EER), and also Receiver Operator Characteristic (ROC) curves. Details for all measures are given in Section 4.1 and Section 5. It is worth underlining that MICHE-I was especially focused on iris segmentation. For this reason more metrics are used to measure performance in this operation. Also the results of proposals addressing iris recognition are discussed with a special consideration for the segmentation methods allowing the best separation of eye regions, and therefore a more reliable feature extraction and matching.

The paper proceeds as follows: Section 2 introduces MICHE-I dataset, that was used both for the challenge and for the further tests presented here. Section 3 summarizes the methods that participated in MICHE-I challenge. Section 4 presents results related to segmentation, and further analyses both the statistical significance of the different performance (PRATT index) of each method between INDOOR and OUTDOOR conditions, and the characteristics of best and worst iris images in terms of achieved segmentation accuracy. Section 5 is devoted to recognition results, mostly in terms of re-combination of the different segmentation and recognition methods, and presents some deeper observations on the key aspects that can affect a good or bad recognition result. Section 5 also sketches time complexity of the different methods. Section 6 presents the tests carried out on the multibiometric combination, and the fusion of results obtained by combining different subsets of the proposed methods. Finally Section 7 draws some conclusions.

2. The MICHE-I database

The aim of the MICHE-I challenge was to assess in a formal and comprehensive way the levels of performance that can be realistically expected from a solution for iris biometrics working on handheld devices. This work provides a deeper insight into its results.

The iris is undoubtedly one of the most popular biometric traits, together with fingerprints and face. It is not unfailing since it can change over time [27], but it is also one of the most reliable biometric trait for a robust recognition. Iris recognition systems have been in fact successfully deployed in various security applications (e.g., airport check and refugee control). However, most of these systems still require that subjects stand close to the capture device (about 1 m or less) and firmly look towards it for a period of about 3 s.

The Chinese Academy of Science collected and made available the first public iris image dataset, named CASIA-Iris, that has been updated from CASIA-IrisV1 to CASIA-IrisV4 since 2002. Its images are collected under near-infrared (NIR) illumination or synthesized. For these reasons, they cannot be reliably used for assessing methods entailing acquisition on mobile devices. In fact, except for a limited percentage of advanced models, these are still mostly equipped with a common RBG camera. The first iris biometric competitions have relied on NIR images as well, that met the mentioned constraints relating to controlled acquisition. Among the most well-known, the Iris Challenge Evaluation (ICE) (http://www.iris.nist.gov/ICE/ , [17]) is worth highlighting, even though the used images share most of the CASIA features and do not represent the type of data expected in mobile environments. Proença and Alexandre [20] have rather tackled the problem of noisy iris recognition. The Noisy Iris Challenge Evaluation (NICE I) they organized exploited images captured in less constrained imaging environments, to evaluate how noise affects iris segmentation (http://www.nice1.di.ubi.pt). To this aim, the proposed iris dataset, namely UBIRIS.v2 [19], contains data captured in visible wavelength (VW), at-a-distance (between 4 and 8 m), and on the move. The results observed confirmed the major impact of uncontrolled conditions on recognition performance. Recognition of VW iris images captured at-a-distance and on the move with less controlled protocols was also the target of the further NICE II contest [21]. Though UBIRIS datasets were captured in visible light and uncontrolled conditions, acquisition was carried out by cameras with much higher resolution than of the images acquired by mobile devices.

In terms of wavelength, note that VW data might contain more minutiae than NIR data (particularly in case of light pigmented irises), but are also much more seriously affected by noisy artifacts (specular reflections) [11]. This raises an apparent contradiction, since more detail does not necessarily means an advantage offered by VW images vs IR ones. On one hand, Hollingsworth et al. [11] noted that humans can recognize more easily images of the periocular region acquired in visible light, since these images show melanin-related differences that do not appear in near-infrared images. On the other hand, however, when tackling iris

[2] The dataset is available on demand at http://www.biplab.unisa.it/MICHE/database/)

recognition, the situation is often reversed. The more clean and easily distinguishable collection of features in NIR images makes the iris recognition problem in NIR generally more feasible than in VW. In addition, performance on NIR images is almost independent of the iris color and pigmentation, while on VW images even a dark pigmentation represents a harder condition compared to the lighter iris colors. Notwithstanding this, some specific patterns are still better detectable in VW. Hosseini et al. [12] discussed how pigment melanin provides a rich feature source in VW, which is unavailable in NIR imaging. This is because, compared to VL [Visible Light], NIR eliminates most of the related information in pigment melanin that scatters in the iris. This is due to the chromophore of the human iris, which has two distinct heterogeneous macromolecules called brown-black Eumelanin and yellow-reddish Pheomelanin. Studying the excitation-emission quantum yields of eumelanin shows that exciting this macromolecule under NIR firing leads to almost no emission of quantum yields where the related chromophors attenuate in NIR imaging [12]. Comparing the advantages of VW/NIR wavelengths, the literature clearly supports the NIR setting, which has induced all commercial iris recognition systems to rely on it. As mobile biometrics are becoming more popular, also NIR attachments become available for mobile devices to implement iris recognition. However, in everyday devices, both NIR sensors and NIR attachments are still quite rare. As mentioned above, the aim of MICHE-I competition, was to assess which level of performance can be achieved without special equipment.

When using a mobile device, it is generally assumed that the subject to be recognized holds and controls the capturing device by himself, though being not necessarily habituated to the data acquisition protocol, and lacking the technical experience to evaluate the capture quality. Note that a more controlled capture would substantially increase the average quality of the data acquired, but would also reduce the challenging levels of the contest, that aims at reproducing real-life as faithfully as possible. Two opposite considerations hold for MICHE-I data: from one side, capturing results might be enhanced by the usually short distance (the length of a human arm, at most), and by the fact that the user tends to assume a frontal pose quite naturally; from the opposite side, the quality of the acquired images suffers from a number of factors: the embedded camera has possibly low resolution, and motion blur, incorrect framing and illumination distortions are also highly probable. These issues call for robust detection/segmentation and encoding procedures. It is worth noting that the accuracy of the latter is heavily affected by the quality of the former. In this context, the MICHE-I dataset represents the starting core of a wider benchmark to be collected thanks to a crowdsourcing approach. This should better allow unbiased assessment of cross-demographic robustness, as well as the interoperability of recognition procedures. In particular, as images are acquired by various mobile devices, the current dataset allows to perceive the cross-sensor recognition effectiveness.

In summary, the key features of MICHE-I dataset are a sufficient population of users, the use of different mobile devices for the collection, the realistic simulation of the acquisition process including different sources of noise, and several acquisition sessions separated in time. A full metadata annotation completes the dataset.

Actually, other mobile datasets published later than MICHE-I include a higher number of subjects. In MICHE-I challenge, the main aim was to investigate the factors negatively affecting iris recognition when capture is carried out using mobile devices. In this context, the relatively lower number of subjects is compensated for by the total number of more than 3000 images which are acquired by different mobile devices, in different conditions. This allows matching samples of the same subject acquired in realistic settings, and to estimate the possible performance degradation. It is worth underlining that cross-device matching is often neglected in literature. Moreover, the number is sufficient to carry out a thorough comparison of different combinations of segmentation/recognition approaches.

The subjects involved in data collection were asked to behave as they would do by using a real system, e.g., subjects wearing eyeglasses could either choose to remove or keep them. They had to take self-images of their iris, by holding the mobile device by themselves, and without any cue about the correctness of iris framing and possible blur. A minimum of four shots for each camera was requested (two out of three devices were equipped with two cameras with different resolutions) and acquisition mode (indoor, outdoor). Indoor acquisition was affected by various sources of artificial light, sometimes combined with natural light ones. Outdoor acquisition was carried out using natural light only. For each subject only one of the two irises was acquired. Three kinds of smartphones/tablets were used for data collection, with Android or Apple iOS operating systems:

- Galaxy Samsung IV (GS4): Google Android; CMOS posterior camera, 13 Megapixels (72 dpi); CMOS anterior camera, 2 Megapixels (72 dpi);
- iPhone5 (IP5): Apple iOS; iSight posterior camera, 8 Megapixels (72 dpi); anterior FaceTime HD Camera, 1.2 Megapixels (72 dpi);
- Galaxy Tablet II (GT2): Google Android; no posterior camera; 0.3 Megapixels anterior camera.

Images have one of three different resolutions (1, 536×2, 048 pixels for iPhone5, 2322×4128 for Galaxy S4, and 640×480 for the tablet). The sources of noise in the MICHE-I dataset include: (a) reflections caused by artificial light sources, natural light sources, people or objects in the scene during the acquisition; (b) focus; (c) blur, either due to an involuntary movement of the hand holding the device, or due to an involuntary movement of the head or of the eye during acquisition; (d) occlusions, due to eyelids, eyeglasses, eyelashes, hair, shadows; (e) device-specific artifacts, due to the low resolution and/or to the specific noise of the device; (f) off-axis gaze; (g) variable illumination; and (h) different color dominants. It is possible to further observe that the lack of precise framing and fixed distance in the capture (both well centered eyes and half faces are present in dataset images), result in variable sizes of the region useful for recognition. This is typical of mobile captures performed by the users, which are usually neither too close nor at arm-length. This introduces further difficulties, since eye localization must be performed in a pre-processing step. In some cases, the resulting size of the iris region is too small, while in other cases it is also possible to exploit the possibilities offered by an extended periocular region. MICHE-I is a multisession dataset, and the time elapsed between the first and second acquisition of a subject varies from a minimum of 2 months to a maximum of 9. At present, MICHE-I contains images from 75 different subjects, with 1297 images from GS4, 1262 images from IP5, and 632 images from GT2.

The Extensible Markup Language (XML) meta-data includes the following tags:

- *filename*: the name of the annotated image; it is composed so to code a certain amount of information in order to quickly find the desired image(s);
- *img_type*: the trait captured in the of image, since face images will be included soon in the dataset;
- *iris*: which iris was acquired (right, left or both);
- *distance_from_the_device*: distance of the user from the acquisition camera, measured to provide a further information for assessment ;
- *session_number*: the number of the acquisition session;
- *image_number*: image ordinal number;

- *user*: id number, age, gender and ethnicity of the subject;
- *device*: all information about the capture device: type, name, camera position (front or rear), resolution and dpi;
- *condition*: information about capture conditions: location, illumination;
- *author*: the name of the laboratory/institution who made that acquisition.

The XML file structure allows a quick and reliable retrieval of any image as a function of any one of the above parameters.

3. Methods participating in the MICHE-I challenge

As discussed above, the MICHE-I dataset contains images acquired in unconstrained settings. Therefore, their average quality is poor, so that the main goal of the participating approaches was to attempt to address such data degradation. To this aim, they mostly used: 1) the periocular region as an extra source of information; 2) color compensation strategies to attenuate the typical difference of sensor features across different devices; and 3) multiple strategies to avoid relying exclusively on a single family of features/methods, therefore reducing the sensitivity to any particular data covariate.

The method proposed by Santos et al. [24] uses both the information from the iris and the periocular region, encoded/matched in a non-holistic way. The idea is to start by segmenting the iris ring, which is also used to define the periocular region-of-interest (ROI). Next, a family of texture descriptors is used to encode the discriminating information in the iris ring and in the regions surrounding the cornea (i.e., eyelids, eyelashes, skin and eyebrows). In more detail, as for periocular region, two types of analysis are applied to the identified ROI: a distribution-based analysis of patches over a grid, and a global analysis of the whole region. The distribution-based analysis involves the computation of local binary patterns (LBP) and histogram of oriented gradients (HOG), and Uniform LBP (ULBP). Each descriptor is computed sequentially for each patch and quantized into histograms. As for global analysis, feature extraction techniques are applied to the whole ROI and the descriptors applied are scale-invariant feature transform (SIFT) and GIST (a set of five scene descriptors [15]). As for iris, information is encoded based on the approach described by Daugman [5] Finally, scores from all the adopted descriptors are fused by a non-linear supervised neural network. Furthermore, the method entails the use of device-specific calibration techniques, that compensate for the different color rendering characterizing each experimental setup. Looking at the results obtained for the contest, it seems that also the latter is one of the keys for such good performance, particularly in the cross-sensor set of tests.

Barra et al. [2] design a complete approach to iris recognition, including segmentation and recognition. The segmentation method, named IS_IS, was originally proposed in [6], and it is modified to run on mobile devices. Segmentation relies on the homogeneity of gray scale histograms of image patches to find the pupil boundary, and on dark-to-light transitions to find the sclera boundary, in a scheme that resembles the well-known Daugman's integro-differential operator. Feature encoding relies on spatial histograms (spatiograms)[3], that can be considered as higher order histograms, that also record the information relating to the spatial domain. They are matched by correlation-based techniques.

The approach by Abate et al. [1] relies on the observation that the features of images from the ocular region acquired by mobile devices are evidently different from the type of data that is generally obtained in more constrained setups. The authors propose an algorithm based on the watershed transform for iris segmentation [30], namely watershed Based IRis Detection (BIRD). The idea is to start by obtaining the gradients in a colored, illumination-corrected image. The final gradient image is obtained by averaging gradients computed over the three channels. Then, the watershed transform is obtained by adopting the topographical distance approach [23]. Next, the output of the watershed transform is used as a guide to binarize the original image and feed a circle detection step, for parametrizing both the pupil and the sclera boundaries. As in several of the competing approaches submitted for MICHE-I, the periocular region is also considered, which is localized using as reference the length of the iris radius. Feature encoding is done by means of 64-bit color histograms, matched using the cosine dissimilarity and Hamming distance.

The idea of Hu et al. [13] is to fuse different previously published iris segmentation techniques, selected according to their performance in particular cases of degraded images. They describe a model selection strategy, which selects the final iris parametrizations based on the candidates returned by the used baseline segmentation strategies. This selection is made according to the image description provided by histograms of local gradients, that are inputted to a support vector machine providing the fused response. This strategy can be easily updated by adding/substituting baseline segmentation methods, and this is an obvious strength of the approach.

The proposal by Haindl and Krupicka [10] is centered on the detection of the non-iris components for the parametrizations of the iris ring. In literature, it is well recognized that the accurate detection of eyelids and reflections is the prerequisite for the accurate iris recognition, both in NIR or VW. The proposed model therefore adaptively learns its parameters on the iris texture part, and subsequently checks for iris reflections using the recursive prediction analysis. After detecting reflections, form-fitting techniques allow finding a parametrization of the pupil. Next, data is converted into the polar domain, where a texture analysis phase is carried out to determine the regions of the normalized data that should not belong to the iris, according to a Bayesian paradigm.

Two methods submitted for MICHE-I can be considered as complementary to the segmentation techniques. Gragnaniello et al. [9] propose an iris liveness detection algorithm for mobile devices. The most innovative point is to use the well known LBP texture descriptor scheme exclusively for the high frequency components of data, which is expected to improve the live/fake discriminability, when compared to the traditional use of this texture descriptor. Bruni and Vitulano [4] propose an application of the modified kernel object tracking to the specific problem of iris tracking. They rely on visual features of human irises that are instinctively used by human eye in the recognition process. Such features are used in the definition of a target feature space as well as of a proposed a metric that well correlates with the way human vision processes and compares information. As a main result, authors argue that one iteration of the mean shift algorithm is enough to get a faithful estimation of iris location in subsequent frames.

Since this paper is particularly focused in presenting the major challenges of iris/ocular recognition in mobile environments, it will not further consider the last two proposals, but will rather concentrate on the participating methods that deal with the issues of segmentation and/or iris verification. The rest of the sections will present an exhaustive analysis of the performance achievable by separating and recombining segmentation and recognition algorithms proposed by competitors, aiming at gaining insights into the open issues and the limitations of mobile iris recognition. Since MICHE-I was especially focused on iris segmentation, more metrics are used to measure performance for this operation. Furthermore, the results of proposals addressing iris recognition are discussed taking into special account the segmentation methods allowing the best separation of eye regions, and therefore a more reliable feature extraction and matching.

Table 1
Metrics used to evaluate the quality of iris segmentation as compared with the manually determined ground truth data.

ACCURACY	Accuracy measures the proportion of **true results** (both true positives and true negatives) with respect to the total number of cases examined.
PRECISION	Precision measures the proportion of the true positives against all the positive results (both true positives and false positives).
SENSITIVITY	Sensitivity is also called the **true positive rate**, or the **recall**, and measures the proportion of positives that are correctly identified as such.
SPECIFICITY	Specificity is also called the **true negative rate**, and measures the proportion of negatives that are correctly identified as such.
F1_SCORE	F1 score is a measure of a test accuracy; it is required to considers both the precision p and the recall r of the test to compute the score; it can be interpreted as a weighted average of the precision and recall, defined as: $F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ F1 score reaches its best value at 1 and worst at 0.
RI	Rand Index counts the fraction of pairs of pixels whose labeling is consistent between the computed segmentation and the ground truth, i.e., the fraction of pairs whose elements are both labeled as edge or as non-edge, both in ground truth and segmentation.
E1_SCORE	The classification error rate (E1) of the algorithm on the input image is given by the proportion of correspondent disagreeing pixels (through the logical EXCLUSIVE-OR operator) over the whole image.
PRATT	This metric is formulated [18] as a function of the distance between correct and measured edge positions, but it is also indirectly related to the false positive and false negative edges; it is defined as: $PRATT = \frac{1}{\max E_G, E_D} \sum_{i=1}^{E_D} \frac{1}{1+\alpha+d_i^2}$ (1) where E_G and E_D are the number of ground truth and detected edge points respectively, d_i is the distance from the *i-th* detected point and the closest ground truth one, and α is a scaling constant set as $\alpha = \frac{1}{9}$ as in the original formulation; the metric reflects the overall behaviour of the distances between the edges, and varies in the range [0,1], where 1 represents the optimal value, i.e., the edges detected coincide with the ground truth.
GCE	The Global Consistency Error (GCE) measures the extent to which one segmentation can be viewed as a refinement of the other; segmentations which are related in this manner are considered to be consistent, since they could represent the same natural image segmented at different scales; details on the computation can be found in [14].
PCC	The Pearson Correlation Coefficient is a measure of the linear correlation between two variables X and Y, returning a value between +1 and inclusive, where 1 is total positive correlation, 0 is no correlation, and is total negative correlation.

4. Segmentation results on MICHE-I DB

All methods have been tested on a subset of MICHE-I that consisted in 591 images for GS4, 571 images for IP5, and 295 images for GT2. For this subset, sequestered ground truth data was created by hand, by manually locating the coordinates of pupil and iris, and the lengths of the radii of the corresponding circumferences. Candidates had no access to this data.

4.1. The metrics

Table 1 summarizes the metrics, typical of binary classification tests, that have been used to evaluate the segmentation quality achieved by the methods submitted to MICHE-I. In practice, they measure the quality of the detected iris contour.

Actually, each metric captures some specific ability of the analyzed algorithms. In general, the ability to correctly classify an existing edge pixel (true positives vs. false negatives) plays a role which is different from the ability to avoid adding false positives (vs. true negatives). They are not symmetrical from the point of view of segmentation algorithms, in the sense that an algorithm that achieves the former might not be so effective to achieve the latter, as it generally happens with binary classifiers. Also the consequences may have a different weight, since a lacking group of pixels in a contour may cause difficulties in detecting a specific contour shape, or produce an unconnected contour where a connected one is needed/expected. The first 4 metrics precisely measure these different aspects separately, to provide a detailed understanding of the positive/negative aspects of each algorithm. On the one hand, *F*1 score is a popular way to get a weighted average of the precision and recall (second and third metrics) in order to have an overall estimate of the ability of the algorithm to distinguish true edge pixels from false ones without missing too many of them. On the other hand *RI* measures the overall agreement between positive/negative classifications and ground truth. $E1_Score$ is in some sense its complement, i.e., the proportion of disagreeing pixels. *PRATT* metric evaluates accuracy from a different point of view, by returning an overall estimate of the actual distance between the detected contours and the ground truth (therefore, in a sense, not only true/false, but also how far from true) and is more specific of segmentation. *GCE* and *PCC* are in a sense higher level measures. *GCE* measures at which extent the errors w.r.t. ground truth can represent a kind of loss of detail in a multi-resolution perspective: errors overall result in a less detailed segmentation though bringing much the same core information. *PCC* is the usual Pearson correlation, to evaluate if the result and the ground truth reflect a similar trend.

4.2. Segmentation Results

Table 2summarizes the scores achieved by participant methods according to the metrics presented in Table 1 to evaluate the segmentation accuracy. It is worth underlining that not all participants presented methods for both segmentation and recognition. Therefore, some papers might be cited only in this section, while others appear only in the next one. Those presenting a complete capture-to-recognition workflow appear in both. For each method and for each device, the first column reports the number of irises that were actually segmented from images captured indoor (IN), outdoor (OUT), and the total; the difference with the original size of the corresponding dataset represents the number of images that were discarded because the segmentation produced null results or threw an exception. The method by Haindl et al. achieves the highest rate of successfully segmented images, while the method used by Barra et al. achieves the lowest. On the other hand, it appears that the usable segmentation results achieved by the latter, although less, are more accurate. In fact, they provide the highest level of similarity with the ground truth (reasonably due to the higher thresholds of acceptance). From the point of view of similarity with ground truth, the second method achieving the best results is the one by Abate et al. The total numbers in absolute suggest that the methods adopted by Barra et al. and Abate et al. provide higher quality masks. Notwithstanding this, the methods by Haindl et al. and Yang et al. are more reliable in terms of rate of success in the following recognition step. These are the reasons that led to a more careful investigation entailing the comparison of the 50 best common segmentations.

A further note that is worth adding is that in Table 2 there is no distinction between front and rear cameras of the devices. As a matter of fact, for this group of experiments, in order to further stress the segmentation algorithms, all images underwent a down-sampling procedure reducing their resolution, therefore mostly canceling the initial device advantage.

Table 2
Comparison of segmentation quality achieved by participant methods according to the measures in Table 1.

			# im.	Acc.	Prec.	Sens.	Spec.	F1	RI	E1	PRATT	GCE	PCC
Barra et al. IS_IS	GS4	IN	196	0.96	0.78	0.80	0.98	0.80	0.92	0.04	0.74	0.05	0.77
		OUT	223	0.96	0.81	0.80	0.98	0.82	0.93	0.04	0.76	0.05	0.78
		Tot	419	0.96	0.80	0.80	0.98	0.81	0.93	0.04	0.75	0.05	0.78
	IP5	IN	225	0.96	0.80	0.81	0.98	0.81	0.92	0.04	0.75	0.05	0.78
		OUT	231	0.96	0.81	0.78	0.98	0.81	0.93	0.04	0.75	0.05	0.77
		Tot	456	0.96	0.80	0.79	0.98	0.81	0.92	0.04	0.75	0.05	0.77
	GT2	IN	66	0.92	0.67	0.74	0.95	0.70	0.87	0.07	0.65	0.08	0.67
		OUT	90	0.96	0.79	0.87	0.97	0.85	0.93	0.04	0.81	0.05	0.82
		Tot	156	0.95	0.69	0.81	0.97	0.79	0.91	0.05	0.74	0.60	0.73
Abate et al. BIRD	GS4	IN	207	0.90	0.53	0.60	0.94	0.54	0.84	0.10	0.52	0.09	0.51
		OUT	210	0.95	0.74	0.78	0.97	0.76	0.91	0.05	0.74	0.06	0.73
		Tot	417	0.92	0.64	0.64	0.95	0.66	0.87	0.07	0.63	0.08	0.62
	IP5	IN	214	0.90	0.60	0.66	0.93	0.59	0.84	0.10	0.58	0.09	0.57
		OUT	226	0.93	0.71	0.72	0.96	0.70	0.88	0.07	0.67	0.07	0.67
		Tot	440	0.92	0.65	0.70	0.94	0.64	0.86	0.08	0.63	0.08	0.62
	GT2	IN	105	0.91	0.63	0.66	0.94	0.60	0.85	0.09	0.60	0.09	0.58
		OUT	125	0.94	0.76	0.72	0.97	0.72	0.90	0.06	0.71	0.07	0.70
		Tot	230	0.93	0.70	0.69	0.96	0.66	0.88	0.07	0.66	0.07	0.65
Haindl et al.	GS4	IN	296	0.94	0.80	0.56	0.98	0.65	0.89	0.06	0.56	0.06	0.63
		OUT	296	0.94	0.89	0.57	0.99	0.69	0.90	0.06	0.58	0.06	0.68
		Tot	591	0.94	0.85	0.57	0.99	0.67	0.89	0.06	0.57	0.06	0.66
	IP5	IN	283	0.94	0.84	0.58	0.99	0.69	0.90	0.06	0.59	0.06	0.67
		OUT	288	0.95	0.92	0.58	0.99	0.71	0.90	0.05	0.59	0.06	0.70
		Tot	571	0.95	0.88	0.58	0.99	0.70	0.90	0.05	0.59	0.06	0.69
	GT2	IN	148	0.94	0.80	0.59	0.98	0.66	0.89	0.06	0.59	0.07	0.65
		OUT	145	0.95	0.94	0.61	0.99	0.74	0.91	0.05	0.62	0.05	0.73
		Tot	293	0.95	0.87	0.60	0.99	0.70	0.90	0.05	0.60	0.06	0.69
Yang et al.	GS4	IN	290	0.95	0.95	0.53	0.99	0.68	0.90	0.05	0.53	0.05	0.68
		OUT	286	0.95	0.96	0.56	0.99	0.73	0.90	0.05	0.56	0.05	0.71
		Tot	576	0.95	0.96	0.55	0.99	0.71	0.90	0.05	0.55	0.05	0.69
	IP5	IN	276	0.95	0.97	0.52	0.99	0.68	0.90	0.05	0.53	0.05	0.68
		OUT	278	0.95	0.97	0.56	0.99	0.72	0.90	0.05	0.57	0.05	0.71
		Tot	554	0.95	0.97	0.54	0.99	0.70	0.90	0.05	0.55	0.05	0.70
	GT2	IN	138	0.94	0.91	0.54	0.99	0.69	0.89	0.06	0.55	0.06	0.67
		OUT	144	0.95	0.95	0.57	0.99	0.72	0.90	0.05	0.57	0.05	0.71
		Tot	282	0.95	0.93	0.55	0.99	0.71	0.90	0.05	0.56	0.06	0.69

4.3. Best 50 common segmentations

This test was carried out to provide a fairer comparison between the methods. As a matter of fact, some of them are based on quality thresholds that cause to discard images where the iris is occluded or not clearly visible. Others (i.e,. Haindl et al. as well as Yang et al.) try to segment everything can be recognized as an iris in the frame. This produces a significantly higher number of segmented irises but, on the negative side, it increments the number of false positives. Considering the issue even from a slightly different point of view, it is possible to observe that the different strategies to identify candidate circumferences are affected by the quality of the acquisition. When processing an adverse image, some segmentation algorithms need much more time to search for a candidate area to be recognized as an iris, and sometimes they do not find it at all.

The images leading to the 50 best segmentations for different subjects, and common for all methods, have been used as a dataset for a new run of comparisons. Table 3 allows observing that, when working on the pictures where the segmentation task is easier, the results are different from those of Table 2 above. Observing the mean scores on all devices, the method by Haindl et al. can be considered the most reliable one, as further testified by examples in Fig. 1.

Due to different and unpredictable behaviour of methods over problematic samples, and in order to establish a common ground for comparison, we report the processing times only for the above "best" samples, i.e., samples that are processed by all methods in a reasonable time. It is implicit that the slowest methods are also those that would encounter the greatest difficulties with adverse samples. Concerning the processing speed, all methods have been tested on an iris section of the image of about 400×300 pixels resolution. The exploited computer is an Intel Xeon X5482 CPU 3,20GH (dual core) 64bit, 10GB RAM.

The segmentations by Barra et al. and Abate et al. overall provide a good compromise between processing speed and segmentation accuracy, as they take (on average) less than 2 seconds for a "good" image. This result is a very positive feature for real time processing. On the other side, the method by Haindl et al. achieves an average segmentation time of about 15 seconds, thus representing a non feasible solution to mobile platforms. Yang et al. is the most time consuming method, with an average time of 35 seconds and more.

5. Recognition Results on MICHE-I

The results given in this section summarize the performance achieved by the recognition methods submitted to MICHE-I in terms of decidability, area under curve (AUC) and equal error rate (EER). The preliminary identification of the iris ROI is carried out using the segmentation methods proposed for the benchmark database.

Decidability is the same FoM used for the NICE II competition [21]. It is obtained by first carrying out a "one-against-all" comparison for each image $I = I_1, \ldots, I_n$ of the data set. The matching process exploits the corresponding binary maps $M = M_1, \ldots, M_n$ that provide the noise-free iris region identified by the segmentation step. This thorough comparison allows to obtain a set of intra-class dissimilarity values $D^I = D^I_1, \ldots, D^I_k$ and a set of inter-class dissimilarity values $D^E = D^E_1, \ldots, D^E_m$, according to whether the pair of images is from the same or from different irises. The decidability

Table 3
Performance measures recomputed considering only the best 50 common segmentations.

Method	Device	PRATT	F1_score	RI	E1_score	GCE	PearsonCC
Barra et al.-IS_IS	GS4	0.962	0.948	0.955	0.024	0.023	0.849
	IP5	0.968	0.961	0.965	0.020	0.020	0.919
	GT2	0.958	0.947	0.958	0.027	0.026	0.855
	MEAN	0.962	0.952	0.960	0.024	0.023	0.874
Abate et al.-BIRD	GS4	0.963	0.966	0.959	0.026	0.021	0.833
	IP5	0.968	0.973	0.959	0.021	0.027	0.865
	GT2	0.974	0.980	0.949	0.026	0.031	0.813
	MEAN	0.968	0.973	0.956	0.026	0.025	0.837
Haindl et al.	GS4	0.980	0.981	0.964	0.019	0.022	0.881
	IP5	0.983	0.986	0.967	0.017	0.029	0.898
	GT2	0.985	0.987	0.954	0.024	0.028	0.822
	MEAN	0.983	0.985	0.961	0.020	0.026	0.867
Yang et al.	GS4	0.986	0.982	0.959	0.020	0.025	0.869
	IP5	0.987	0.988	0.952	0.020	0.027	0.849
	GT2	0.984	0.987	0.962	0.019	0.021	0.862
	MEAN	0.986	0.986	0.958	0.020	0.024	0.860

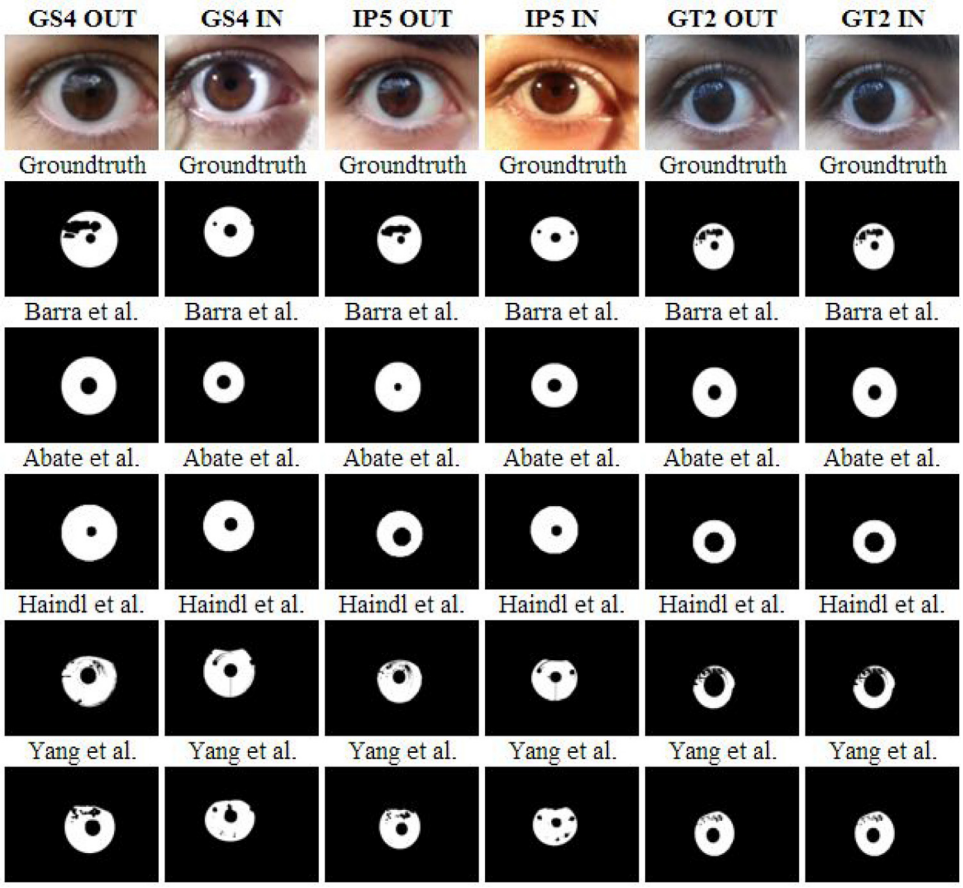

Fig. 1. An example of images resulting in a good segmentation common to all methods.

value $d'(D^I_1, \ldots, D^I_k, D^E_1, \ldots, D^E_m) \rightarrow [0, \infty[$ used as evaluation measure is computed as:

$$d' = \frac{|avg(D^I) - avg(D^E)|}{\sqrt{\frac{1}{2} \times (\sigma^2(D^I) + \sigma^2(D^E))}}, \tag{2}$$

where $avg(D^I)$ and $avg(D^E)$ denote the average values of the intra-class and inter-class comparisons and $\sigma^2(D^I)$ and $\sigma^2(D^E)$ are the corresponding variance values. We implemented a "bootstrapping-like" approach for the computation of 95% Confidence Intervals (CI_low CI_high) an d the Bootstrapped Standard Error (SE) for DEC, EER and AUC associated to each experiment. The measurements are reported in the form: $VALUE \pm SE$ (CI_low CI_high).

The recognition algorithms considered are presented in [1,2,22], and [24]. As anticipated, when a same proposal contained both a segmentation and a recognition algorithm, and these were clearly separable, they were extracted and recombined in all possible ways. In other words, whether the participants used a segmentation algorithm of their own of not, we tested all recognition methods with all segmentation ones. It is worth pointing out that, when recognition is assessed, it is of great interest to also test cross-device performance. To this aim, sets of images acquired by the same device are alternatively used as either probe set (gallery) or as test set (or as both, for intra-device recognition). All possible combinations of probe/gallery images were included in the tests.

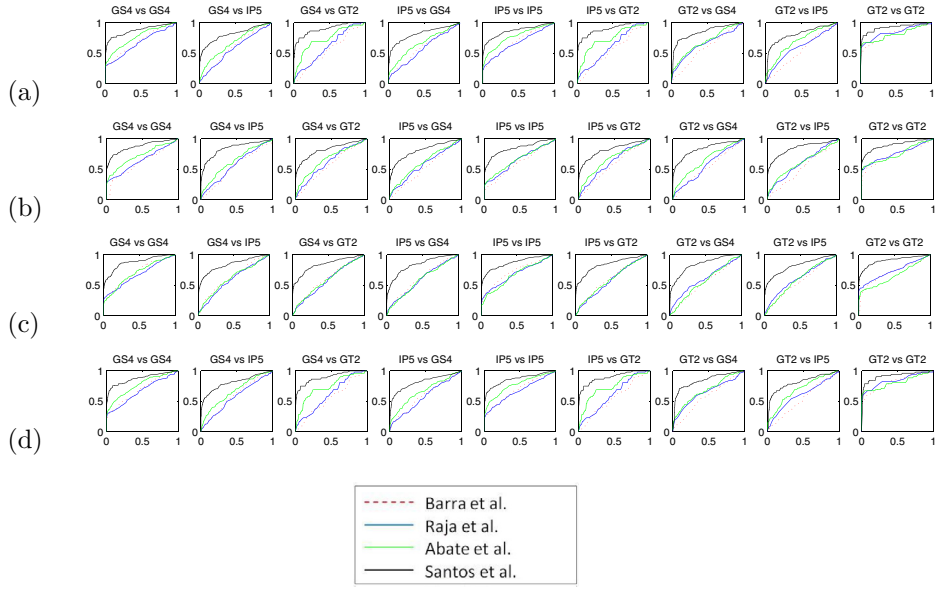

Fig. 2. ROC curves for each pair of devices (FAR on horizontal axis and GAR on vertical axis). The figure is divided into four areas. From top to bottom, row (a) collects the cross-device ROC curves achieved when segmenting with IS_IS algorithm. The second row (b) of plots refers to results achieved when BIRD is used. The third row (c) shows the results achieved when segmenting by Haindl et al. algorithm, and the last one (d) is related to the use of the segmentation algorithm by Yang et al.

Since the main target of MICHE-I was segmentation, results will be presented and discussed taking it as the principal variable element. Tables 4 and 5 report the results achieved for each segmentation algorithm, considering the different recognition algorithms and the different intra- and inter-device classes of comparison. Also, Fig. 2 presents an overall view of the ROC curves related to the same test conditions, to provide a more immediate insight on the levels of interoperability (cross-sensor recognition) supported by each method. To this aim, curves relating to different recognition algorithms, tested on the same combination of probe/gallery, after the same segmentation, are grouped together. In all cases, it is obvious how significantly the recognition method by Santos et al. outperforms the other three in all classes of comparisons and with all segmentation algorithms. Moreover, it is possible to observe that the class of comparison GT2vsGT2 achieves a higher level of performance (on average and compared to the others) in terms of ROC, notwithstanding the poorer resolution of the embedded camera. This is confirmed by inspecting the EER values in the tables, even though this class generally presents the lowest decidability values. However, these results do not take into account that the sizes of probe and gallery sets for GT2 are smaller than the others. The lower percentage of uncertainty contributes to increase the level of performance. As expected, it is possible to notice a general decrease of performance in cross-device operations, except for some cases, that will be underlined when appropriate. It is also interesting to notice that performances are generally affected by the exchange in the probe/gallery role of images taken by the different devices. In the following subsections, we organize the recognition results by the exploited segmentation algorithm.

Results by segmenting with the IS_IS mobile algorithm

Table 4 (top part) summarizes the performance in terms of the decidability, AUC, and EER, of the different recognition methods when segmenting with IS_IS algorithm. It should be noted that all methods achieved better performance, except for decidability, on GT2vsGT2. Fig. 2 (a) allows to better appreciate both the higher recognition performance of Santos et al., and the good behaviour of GT2vsGT2 class of test.

Results by segmenting with the BIRD algorithm

Table 4 (bottom part) shows the performance of the different recognition methods in terms of decidability, AUC, and EER, when exploiting Abate et al. algorithm for segmentation (BIRD, based on watershed). Both the bottom part of Table 4 and Fig. 2 (b) confirm the same general trend underlined for IS_IS segmentation: Santos et al. is the method providing the best recognition performance, and GT2vsGT2 is the class with the best behaviour. It can be noticed how ROC curves of Santos et al. are definitely better than those for the other methods, that present very similar trends instead.

Results by segmenting with the algorithm by Haindl et al.

Table 5 (top part) provides the recognition performance of the different recognition methods in terms of decidability, AUC, and EER, when exploiting the segmentation method proposed in Haindl et al. Besides observing an even higher superiority of Santos et al., it is interesting to also notice how, with this segmentation method, the ROC curves of the other recognition algorithms are grouped within the same band, that is even narrower than when exploiting IS_IS and BIRD segmentation methods (see Fig. 2 (c)).

Results by segmenting with the algorithm by Yang et al.

Table 5 (bottom part) shows the achieved performance of the different recognition methods in terms of decidability, AUC, and EER, when exploiting the segmentation method proposed in Yang et al. The method by Yang et al. further accentuates the behavior observed when using Haindl et al. method. It is worth noticing that, with this segmentation, the recognition by Raja et al. is the one achieving the best intra-device recognition in terms of decidability (see also Fig. 2 (d)). In summary, it is possible to observe that these last two segmentation methods are somehow more stable, since they provide results that allow less performance difference in the following recognition step.

5.1. Pairwise score comparison

In order to carry out some deeper observations on the key aspects that can affect either a good or a bad recognition result, the best and worst samples per experiment have been extracted from pairwise comparisons. In this context, an experiment is represented by the combination of a capture device, a segmentation method and a recognition algorithm. In particular, the focus is on the best/worst pairwise comparisons that were common to all experiments. In more detail, for each experiment, each gallery tem-

Table 4
Performance of the recognition methods in terms of decidability, AUC and EER, when segmenting with the algorithms by Barra *et al*(top) and by Abate *et al*(bottom). The best results for each intra-/inter-device test are in bold, those for each method are underlined.

Device	Recognition Method	GS4			IP5			GT2		
		DEC	AUC	EER	DEC	AUC	EER	DEC	AUC	EER
Barra et al. segmentation algorithm (IS_IS)										
GS4	Barra et al.	1.871 ± 0.023 (1.824–1.915)	0.693 ± 0.008 (0.677–0.708)	0.394 ± 0.007 (0.381–0.407)	2.120 ± 0.016 (2.089–2.146)	0.593 ± 0.008 (0.578–0.610)	0.428 ± 0.007 (0.414–0.440)	1.935 ± 0.035 (1.885–2.032)	0.555 ± 0.012 (0.529–0.579)	0.461 ± 0.016 (0.434–0.490)
	Raja et al.	4.972 ± 0.273 (4.760–5.697)	0.664 ± 0.007 (0.650–0.677)	0.400 ± 0.007 (0.385–0.414)	2.731 ± 0.171 (2.377–3.114)	0.569 ± 0.004 (0.561–0.577)	0.449 ± 0.005 (0.439–0.457)	3.312 ± 0.180 (2.963–3.742)	0.612 ± 0.011 (0.590–0.635)	0.433 ± 0.011 (0.411–0.455)
	Abate et al.	**8.203** ± 0.186 (7.569–8.386)	0.764 ± 0.007 (0.752–0.779)	0.322 ± 0.006 (0.309–0.334)	4.433 ± 0.050 (4.309–4.517)	0.665 ± 0.008 (0.650–0.681)	0.386 ± 0.010 (0.369–0.409)	4.172 ± 0.118 (4.032–4.553)	0.739 ± 0.013 (0.716–0.767)	0.307 ± 0.021 (0.261–0.346)
	Santos et al.	6.211 ± 0.180 (5.695–6.483)	**0.874** ± 0.007 (0.864–0.889)	**0.210** ± 0.011 (0.185–0.226)	**5.972** ± 0.112 (5.797–6.214)	**0.811** ± 0.008 (0.798–0.828)	**0.254** ± 0.006 (0.242–0.267)	**5.030** ± 0.311 (4.818–6.038)	**0.897** ± 0.009 (0.886–0.917)	**0.183** ± 0.010 (0.156–0.197)
IP5	Barra et al.	2.192 ± 0.029 (2.138–2.247)	0.590 ± 0.007 (0.576–0.603)	0.430 ± 0.008 (0.414–0.446)	2.341 ± 0.020 (2.302–2.381)	0.692 ± 0.007 (0.679–0.705)	0.374 ± 0.006 (0.362–0.385)	2.387 ± 0.036 (2.325–2.459)	0.529 ± 0.014 (0.500–0.554)	0.467 ± 0.011 (0.445–0.487)
	Raja et al.	3.157 ± 0.181 (2.994–3.725)	0.570 ± 0.005 (0.560–0.580)	0.458 ± 0.008 (0.443–0.473)	5.393 ± 0.299 (5.180–6.378)	0.667 ± 0.006 (0.655–0.678)	0.384 ± 0.008 (0.370–0.398)	3.472 ± 0.224 (3.003––4.055)	0.579 ± 0.013 (0.552–0.604)	0.443 ± 0.013 (0.419–0.470)
	Abate et al.	4.805 ± 0.080 (4.578–4.917)	0.662 ± 0.008 (0.648–0.677)	0.375 ± 0.008 (0.361–0.391)	7.201 ± 0.079 (7.008–7.338)	0.771 ± 0.007 (0.759–0.786)	0.293 ± 0.007 (0.279–0.306)	4.025 ± 0.047 (3.903–4.090)	0.671 ± 0.013 (0.646–0.696)	0.412 ± 0.011 (0.392–0.434)
	Santos et al.	**10.557** ± 1.657 (6.036–11.408)	**0.817** ± 0.009 (0.804–0.837)	**0.255** ± 0.007 (0.240–0.267)	**11.400** ± 1.734 (6.082–12.297)	**0.865** ± 0.006 (0.855–0.878)	**0.217** ± 0.008 (0.202–0.231)	**7.730** ± 1.004 (4.873–9.528)	**0.834** ± 0.012 (0.815–0.860)	**0.212** ± 0.017 (0.179–0.242)
GT2	Barra et al.	2.208 ± 0.034 (2.145–2.275)	0.576 ± 0.010 (0.558–0.596)	0.432 ± 0.015 (0.407–0.463)	2.540 ± 0.027 (2.486–2.591)	0.539 ± 0.010 (0.519–0.558)	0.490 ± 0.007 (0.475–0.503)	1.929 ± 0.030 (1.869–1.984)	0.814 ± 0.010 (0.796–0.836)	0.288 ± 0.016 (0.254–0.314)
	Raja et al.	1.747 ± 0.772 (1.230–3.971)	0.647 ± 0.009 (0.629–0.664)	0.380 ± 0.012 (0.355–0.402)	1.492 ± 0.585 (1.250–3.472)	0.601 ± 0.008 (0.585–0.616)	0.417 ± 0.010 (0.393–0.435)	3.240 ± 1.535 (2.614–7.495)	0.846 ± 0.011 (0.825–0.870)	0.225 ± 0.012 (0.200–0.245)
	Abate et al.	4.204 ± 0.058 (4.088–4.321)	0.665 ± 0.010 (0.647–0.682)	0.376 ± 0.015 (0.347–0.403)	4.447 ± 0.084 (4.229–4.605)	0.674 ± 0.009 (0.658–0.691)	0.355 ± 0.009 (0.339–0.372)	7.531 ± 0.079 (7.061–7.624)	0.792 ± 0.013 (0.771–0.820)	0.289 ± 0.016 (0.262–0.317)
	Santos et al.	**4.871** ± 0.272 (4.660–5.670)	**0.844** ± 0.008 (0.830–0.862)	**0.236** ± 0.008 (0.220–0.250)	**4.910** ± 0.268 (4.701–5.736)	**0.811** ± 0.010 (0.793–0.834)	**0.250** ± 0.011 (0.230–0.269)	**4.207** ± 0.186 (4.024–4.738)	**0.918** ± 0.009 (0.907–0.940)	**0.153** ± 0.016 (0.113–0.177)
Abate et al. segmentation algorithm (BIRD)										
GS4	Barra et al.	1.739 ± 0.031 (1.676–1.796)	0.608 ± 0.008 (0.594–0.625)	0.418 ± 0.007 (0.404–0.431)	2.368 ± 0.020 (2.327–2.406)	0.540 ± 0.008 (0.525–0.558)	0.448 ± 0.005 (0.439–0.458)	1.774 ± 0.029 (1.710–1.830)	0.567 ± 0.008 (0.553–0.582)	0.449 ± 0.008 (0.434–0.465)
	Raja et al.	6.830 ± 0.125 (6.652–7.109)	0.646 ± 0.007 (0.633–0.659)	0.408 ± 0.006 (0.394–0.418)	3.917 ± 0.094 (3.775–4.129)	0.575 ± 0.006 (0.562–0.586)	0.450 ± 0.007 (0.436–0.464)	3.788 ± 0.068 (3.674–3.938)	0.618 ± 0.006 (0.607–0.630)	0.424 ± 0.007 (0.410–0.438)
	Abate et al.	**6.916** ± 0.055 (6.833–7.047)	0.724 ± 0.008 (0.710–0.739)	0.337 ± 0.008 (0.322–0.352)	4.643 ± 0.057 (4.531–4.762)	0.636 ± 0.009 (0.619–0.654)	0.420 ± 0.013 (0.388–0.440)	4.847 ± 0.129 (4.476–4.975)	0.664 ± 0.009 (0.646–0.682)	0.369 ± 0.007 (0.355–0.383)
	Santos et al.	5.761 ± 0.165 (5.572–6.280)	**0.873** ± 0.006 (0.863–0.886)	**0.204** ± 0.007 (0.193–0.219)	**5.527** ± 0.123 (5.346–5.832)	**0.845** ± 0.007 (0.834–0.860)	**0.218** ± 0.005 (0.209–0.227)	**5.506** ± 0.220 (5.302–6.156)	**0.823** ± 0.007 (0.810–0.838)	**0.247** ± 0.009 (0.227–0.261)
IP5	Barra et al.	2.204 ± 0.043 (2.127–2.295)	0.523 ± 0.007 (0.510–0.537)	0.476 ± 0.007 (0.464–0.491)	2.629 ± 0.025 (2.585–2.676)	0.602 ± 0.008 (0.586–0.619)	0.437 ± 0.008 (0.420–0.451)	2.236 ± 0.044 (2.152–2.326)	0.501 ± 0.006 (0.489–0.513)	0.491 ± 0.007 (0.477–0.504)
	Raja et al.	3.225 ± 0.087 (3.095–3.442)	0.575 ± 0.006 (0.564–0.586)	0.450 ± 0.010 (0.428–0.466)	**6.752** ± 0.221 (6.334–7.323)	0.647 ± 0.006 (0.635–0.660)	0.397 ± 0.008 (0.381–0.411)	4.021 ± 0.157 (3.556–4.281)	0.538 ± 0.006 (0.527–0.548)	0.474 ± 0.010 (0.455–0.491)
	Abate et al.	4.771 ± 0.117 (4.481–4.914)	0.613 ± 0.009 (0.596–0.630)	0.429 ± 0.009 (0.410–0.445)	6.765 ± 0.087 (6.676–6.931)	0.661 ± 0.007 (0.647–0.675)	0.391 ± 0.008 (0.378–0.409)	5.290 ± 0.304 (4.391–5.483)	0.578 ± 0.008 (0.562–0.595)	0.448 ± 0.008 (0.432–0.463)
	Santos et al.	**5.650** ± 0.154 (5.470–6.109)	**0.811** ± 0.008 (0.798–0.829)	**0.259** ± 0.007 (0.246–0.272)	5.800 ± 0.147 (5.435–6.056)	**0.846** ± 0.007 (0.836–0.861)	**0.220** ± 0.008 (0.203–0.233)	**10.905** ± 1.433 (6.076–11.656)	**0.805** ± 0.007 (0.793–0.822)	**0.273** ± 0.009 (0.257–0.291)
GT2	Barra et al.	1.698 ± 0.026 (1.650–1.748)	0.540 ± 0.007 (0.525–0.553)	0.478 ± 0.010 (0.462–0.500)	2.344 ± 0.019 (2.305–2.379)	0.519 ± 0.008 (0.503–0.534)	0.474 ± 0.006 (0.463–0.488)	1.883 ± 0.026 (1.830–1.931)	0.742 ± 0.009 (0.725–0.761)	0.315 ± 0.009 (0.299–0.334)
	Raja et al.	4.418 ± 0.222 (3.849–4.854)	0.569 ± 0.006 (0.557–0.581)	0.447 ± 0.007 (0.432–0.461)	4.552 ± 0.155 (4.352–4.952)	0.599 ± 0.006 (0.588–0.611)	0.437 ± 0.007 (0.422–0.448)	**10.011** ± 0.412 (9.611–11.090)	0.761 ± 0.008 (0.746–0.777)	0.323 ± 0.012 (0.294–0.343)
	Abate et al.	4.937 ± 0.054 (4.839–5.057)	0.654 ± 0.008 (0.638–0.670)	0.410 ± 0.006 (0.398–0.422)	4.587 ± 0.109 (4.483–4.903)	0.607 ± 0.008 (0.589–0.624)	0.425 ± 0.009 (0.410–0.443)	8.222 ± 0.072 (8.122–8.402)	0.755 ± 0.009 (0.739–0.775)	0.345 ± 0.006 (0.331–0.357)
	Santos et al.	**5.901** ± 0.106 (5.730–6.157)	**0.839** ± 0.009 (0.826–0.857)	**0.236** ± 0.008 (0.215–0.250)	**5.838** ± 0.111 (5.666–6.060)	**0.845** ± 0.007 (0.833–0.860)	**0.236** ± 0.007 (0.218–0.248)	6.892 ± 0.597 (5.185–7.343)	**0.901** ± 0.007 (0.891–0.915)	**0.165** ± 0.007 (0.153–0.178)

Table 5
Performance of the recognition methods in terms of decidability, AUC and EER, when segmenting with the algorithms by Haindl *et al* (top) and by Yang *et al* (bottom). The best results for each intra-/inter-device test are in bold, those for each method are underlined.

Device	Recognition Method	GS4			IP5			GT2		
		DEC	AUC	EER	DEC	AUC	EER	DEC	AUC	EER
Haindl et al. segmentation algorithm										
GS4	Barra et al.	2.640 ± 0.023 (2.587–2.681)	0.707 ± 0.007 (0.694–0.720)	0.368 ± 0.005 (0.358–0.378)	2.819 ± 0.024 (2.757–2.855)	0.631 ± 0.008 (0.616–0.646)	0.417 ± 0.008 (0.401–0.431)	2.525 ± 0.034 (2.462–2.596)	0.605 ± 0.006 (0.594–0.617)	0.430 ± 0.007 (0.416–0.444)
	Raja et al.	**6.488** ± 0.173 (6.267–6.977)	0.673 ± 0.005 (0.663–0.682)	0.388 ± 0.005 (0.377–0.397)	3.059 ± 0.080 (2.949–3.279)	0.592 ± 0.005 (0.583–0.601)	0.436 ± 0.006 (0.426–0.449)	3.767 ± 0.134 (3.462–4.079)	0.573 ± 0.004 (0.564–0.581)	0.461 ± 0.006 (0.449–0.471)
	Abate et al.	5.355 ± 0.075 (5.244–5.505)	0.699 ± 0.006 (0.687–0.711)	0.358 ± 0.009 (0.343–0.376)	4.469 ± 0.095 (4.256–4.704)	0.610 ± 0.007 (0.598–0.624)	0.422 ± 0.006 (0.408–0.433)	4.961 ± 0.057 (4.843–5.070)	0.583 ± 0.007 (0.570–0.596)	0.451 ± 0.006 (0.439–0.463)
	Santos et al.	6.131 ± 0.143 (5.811–6.400)	**0.878** ± 0.005 (0.869–0.889)	**0.198** ± 0.005 (0.187–0.207)	**6.444** ± 0.191 (5.903–6.750)	**0.840** ± 0.006 (0.829–0.851)	**0.235** ± 0.008 (0.220–0.249)	**6.639** ± 0.217 (6.020–6.950)	**0.840** ± 0.007 (0.828–0.855)	**0.229** ± 0.007 (0.216–0.243)
IP5	Barra et al.	2.907 ± 0.028 (2.853–2.959)	0.615 ± 0.007 (0.602–0.630)	0.429 ± 0.011 (0.407–0.449)	2.917 ± 0.019 (2.876–2.950)	0.718 ± 0.007 (0.705–0.732)	0.339 ± 0.007 (0.325–0.350)	3.010 ± 0.051 (2.908–3.117)	0.567 ± 0.007 (0.554–0.581)	0.452 ± 0.007 (0.436–0.465)
	Raja et al.	3.199 ± 0.261 (2.956–3.937)	0.593 ± 0.004 (0.585–0.601)	0.441 ± 0.007 (0.425–0.455)	**6.207** ± 0.438 (5.927–7.566)	0.677 ± 0.005 (0.667–0.688)	0.386 ± 0.007 (0.374–0.400)	3.678 ± 0.284 (3.134–4.487)	0.590 ± 0.005 (0.579–0.599)	0.417 ± 0.008 (0.404–0.437)
	Abate et al.	4.360 ± 0.063 (4.194–4.469)	0.585 ± 0.008 (0.571–0.601)	0.431 ± 0.007 (0.417–0.444)	5.610 ± 0.056 (5.515–5.710)	0.664 ± 0.007 (0.652–0.678)	0.379 ± 0.006 (0.367–0.389)	4.918 ± 0.069 (4.734–5.006)	0.602 ± 0.007 (0.590–0.615)	0.430 ± 0.006 (0.418–0.443)
	Santos et al.	**5.356** ± 0.079 (5.223–5.516)	**0.838** ± 0.006 (0.827–0.849)	**0.247** ± 0.008 (0.229–0.260)	5.438 ± 0.085 (5.304–5.620)	**0.868** ± 0.006 (0.859–0.880)	**0.220** ± 0.007 (0.206–0.233)	**7.199** ± 0.420 (5.773–7.531)	**0.818** ± 0.007 (0.805–0.833)	**0.265** ± 0.006 (0.251–0.277)
GT2	Barra et al.	2.384 ± 0.018 (2.349–2.418)	0.540 ± 0.009 (0.524–0.558)	0.462 ± 0.007 (0.449–0.474)	2.868 ± 0.016 (2.834–2.898)	0.554 ± 0.009 (0.535–0.571)	0.467 ± 0.010 (0.450–0.485)	2.048 ± 0.024 (1.999–2.094)	0.713 ± 0.006 (0.702–0.727)	0.339 ± 0.006 (0.327–0.350)
	Raja et al.	3.767 ± 0.163 (3.487–4.152)	0.605 ± 0.006 (0.594–0.617)	0.432 ± 0.010 (0.415–0.451)	4.054 ± 0.198 (3.683–4.477)	0.609 ± 0.006 (0.598–0.619)	0.422 ± 0.005 (0.413–0.431)	**8.630** ± 0.414 (8.194–9.713)	0.738 ± 0.003 (0.731–0.745)	0.339 ± 0.004 (0.330–0.347)
	Abate et al.	4.594 ± 0.092 (4.364–4.758)	0.565 ± 0.007 (0.552–0.578)	0.467 ± 0.007 (0.451–0.480)	4.791 ± 0.056 (4.685–4.899)	0.589 ± 0.007 (0.576–0.603)	0.432 ± 0.006 (0.419–0.444)	6.467 ± 0.064 (6.339–6.562)	0.660 ± 0.006 (0.650–0.671)	0.422 ± 0.005 (0.412–0.432)
	Santos et al.	**6.624** ± 0.136 (6.403–6.910)	**0.847** ± 0.006 (0.837–0.859)	**0.224** ± 0.005 (0.214–0.234)	**6.799** ± 0.158 (6.570–7.157)	**0.818** ± 0.007 (0.807–0.832)	**0.270** ± 0.006 (0.257–0.282)	6.201 ± 0.121 (6.012–6.465)	**0.908** ± 0.004 (0.900–0.916)	**0.163** ± 0.007 (0.150–0.176)
Yang et al. segmentation algorithm										
GS4	Barra et al.	1.201 ± 0.023 (1.155–1.240)	0.583 ± 0.006 (0.572–0.596)	0.434 ± 0.007 (0.420–0.448)	1.761 ± 0.016 (1.727–1.787)	0.539 ± 0.008 (0.526–0.554)	0.473 ± 0.006 (0.462–0.484)	1.357 ± 0.022 (1.309–1.398)	0.532 ± 0.006 (0.522–0.543)	0.480 ± 0.007 (0.467–0.493)
	Raja et al.	**6.286** ± 0.290 (6.044–7.254)	0.662 ± 0.005 (0.652–0.673)	0.401 ± 0.008 (0.387–0.416)	3.382 ± 0.068 (3.268–3.522)	0.594 ± 0.004 (0.587–0.601)	0.437 ± 0.006 (0.425–0.448)	3.449 ± 0.103 (3.196–3.663)	0.587 ± 0.005 (0.576–0.597)	0.451 ± 0.007 (0.436–0.464)
	Abate et al.	5.227 ± 0.085 (5.108–5.464)	0.664 ± 0.007 (0.651–0.677)	0.386 ± 0.005 (0.376–0.396)	3.830 ± 0.056 (3.729–3.952)	0.591 ± 0.008 (0.576–0.608)	0.427 ± 0.006 (0.416–0.439)	4.498 ± 0.045 (4.389–4.571)	0.604 ± 0.007 (0.588–0.618)	0.437 ± 0.007 (0.425–0.452)
	Santos et al.	6.077 ± 0.134 (5.893–6.427)	**0.899** ± 0.006 (0.891–0.912)	**0.176** ± 0.006 (0.162–0.185)	**5.875** ± 0.106 (5.721–6.142)	**0.886** ± 0.005 (0.877–0.898)	**0.190** ± 0.005 (0.180–0.198)	**6.396** ± 0.145 (6.064–6.667)	**0.860** ± 0.006 (0.850–0.874)	**0.205** ± 0.005 (0.196–0.214)
IP5	Barra et al.	1.825 ± 0.038 (1.760–1.901)	0.537 ± 0.004 (0.530–0.544)	0.472 ± 0.005 (0.463–0.480)	2.231 ± 0.029 (2.177–2.288)	0.663 ± 0.006 (0.652–0.675)	0.381 ± 0.005 (0.371–0.390)	1.906 ± 0.037 (1.832–1.981)	0.531 ± 0.004 (0.523–0.538)	0.471 ± 0.005 (0.461–0.479)
	Raja et al.	3.537 ± 0.225 (3.132–4.234)	0.605 ± 0.004 (0.597–0.613)	0.443 ± 0.005 (0.433–0.453)	**6.382** ± 0.373 (6.077–7.485)	0.670 ± 0.005 (0.661–0.678)	0.396 ± 0.006 (0.381–0.407)	3.311 ± 0.182 (3.157–3.887)	0.587 ± 0.006 (0.575–0.597)	0.439 ± 0.005 (0.428–0.449)
	Abate et al.	3.955 ± 0.054 (3.864–4.088)	0.648 ± 0.006 (0.636–0.660)	0.391 ± 0.007 (0.378–0.403)	4.744 ± 0.077 (4.637–4.959)	0.684 ± 0.008 (0.671–0.699)	0.390 ± 0.006 (0.375–0.401)	4.533 ± 0.052 (4.392–4.599)	0.558 ± 0.007 (0.544–0.573)	0.456 ± 0.006 (0.445–0.468)
	Santos et al.	**5.823** ± 0.123 (5.665–6.176)	**0.866** ± 0.005 (0.856–0.877)	**0.215** ± 0.005 (0.207–0.228)	5.615 ± 0.085 (5.473–5.816)	**0.887** ± 0.005 (0.878–0.897)	**0.198** ± 0.006 (0.185–0.213)	**6.036** ± 0.132 (5.877–6.389)	**0.838** ± 0.006 (0.828–0.851)	**0.245** ± 0.008 (0.225–0.258)
GT2	Barra et al.	1.281 ± 0.018 (1.246–1.315)	0.538 ± 0.004 (0.531–0.546)	0.472 ± 0.005 (0.463–0.483)	1.834 ± 0.014 (1.806–1.860)	0.552 ± 0.006 (0.541–0.564)	0.466 ± 0.005 (0.455–0.475)	1.429 ± 0.023 (1.390–1.486)	0.685 ± 0.005 (0.677–0.695)	0.358 ± 0.005 (0.349–0.369)
	Raja et al.	5.263 ± 0.411 (4.118–5.659)	0.587 ± 0.004 (0.580–0.594)	0.449 ± 0.006 (0.438–0.460)	3.743 ± 0.124 (3.445–3.973)	0.574 ± 0.004 (0.566–0.581)	0.451 ± 0.006 (0.440–0.464)	**7.227** ± 0.158 (6.995–7.582)	0.736 ± 0.004 (0.729–0.744)	0.334 ± 0.004 (0.327–0.341)
	Abate et al.	4.605 ± 0.103 (4.317–4.724)	0.600 ± 0.006 (0.589–0.612)	0.428 ± 0.006 (0.417–0.439)	4.442 ± 0.053 (4.340–4.545)	0.568 ± 0.005 (0.559–0.579)	0.444 ± 0.004 (0.435–0.453)	5.964 ± 0.083 (5.769–6.052)	0.712 ± 0.005 (0.702–0.723)	0.350 ± 0.007 (0.337–0.363)
	Santos et al.	**6.728** ± 0.216 (6.051–7.038)	**0.858** ± 0.005 (0.849–0.869)	**0.216** ± 0.005 (0.206–0.224)	**6.647** ± 0.206 (6.038–6.935)	**0.855** ± 0.005 (0.847–0.866)	**0.227** ± 0.005 (0.218–0.236)	6.094 ± 0.164 (5.597–6.311)	**0.924** ± 0.004 (0.918–0.932)	**0.143** ± 0.004 (0.136–0.150)

Table 6
Iris samples providing the best and worst recognition results shared across all recognition methods (see the text for the definition of "best" and "worst").

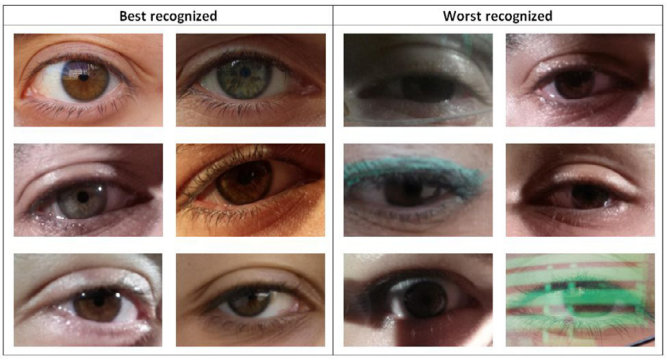

plate is compared with each of the others of the same subject, and the full set of the obtained intra-subject dissimilarity scores is ordered by ascending values. Of course, such scores may fall in different ranges and have different distributions among the methods, therefore the numerical results of different experiments are not always directly comparable. However, it is still possible and interesting to compare the obtained rankings. The samples considered as the "best" ones, always achieve very good similarity when compared with samples of the same subject, and the resulting values appear in the first positions of all the described ordered lists. The symmetric behavior holds for the "worst" samples, when compared with samples of the same subject. In other words, the best samples always achieve low dissimilarity scores for the same subject in probe/gallery (of course for different captures), while worst samples always achieve high dissimilarity scores for the same subject in probe/gallery. Reporting all the results observed for this task, for all possible experiments carried out in this study, would have been relatively difficult and hard to summarise, as well as cumbersome for the reader. Therefore, we selected only the most representative either positive or negative results, that testify either a similar or contradictory trend in all experiments. Out of this subset, we selected the acquisitions which enable to make the most relevant observations regarding their characterizing features, and, consequently, their impact on recognition performances. This provides an easier to analyze and comprehensive point of view on the results, while reducing redundant information.

The images in the right part of Table 6 illustrate some examples of the iris samples that produced the worst recognition results for genuine/impostor pairwise comparisons. By analysing their features, it is possible to observe that the occlusions by the eyelids are rather evident in most of the pictures. Also, the average brightness of images is low, or the iris falls in a region affected by shadows, a condition that makes the extraction of iris features generally hard. The last picture shown at the bottom of the fourth column represents an extremely challenging yet quite common condition in outdoor settings, where a recognition system may fail. Even if its contribution to the scope of this section is rather limited, it is a useful candidate to illustrate the level of complexity of the MICHE-I contest (there are many pictures that present this kind of environmental noise). Conversely, the relevant well recognized subjects also present a collection of non-trivial samples (on the left in Table 6) taken either in outdoor or indoor conditions. From the good results obtained with such samples, it is possible to observe that in "good" samples the visibility of the irises and of the pupils is high, thus making it easier to detect and segment them. Furthermore, differently from expected, the environmental reflections on eye surface in outdoor condition do not necessarily imply a drop

Table 7
List of verification scores (dissimilarities) in the range [0,1] on the iris samples in Fig. 3, that provides unpredictably different results among the recognition methods; the worst result is underlined, the best one is in bold.

Segmentation	Recognition			
	Barra et al.	Raja et al.	Abate et al.	Santos et al.
Barra et al.	0.104	0.356	<u>0.669</u>	0.136
Abate et al.	**0.007**	0.253	0.286	0.207
Haindl et al.	0.455	0.429	0.637	0.636
Yang et al.	0.075	0.457	0.594	0.332

in performances. Therefore, it is interesting to investigate the cases where such reflections rather hinder a reliable processing. Fig. 3 shows, on the left, an interesting typical example of this condition, with an indoor sample of the same subject on the right.

The iris sample on the left of Fig. 3 is interesting because it presents an extremely good level of detail due to the visibility of good iris contours and a satisfactory illumination. Notwithstanding this, it has been selected as a representative candidate of many similar acquisitions in MICHE-I, for which the recognition methods provided discordant results in verification mode, when comparison was carried out with a good template of the same subject. It is possible to observe that in this image the iris region is significantly occluded by environmental reflections. On one hand, this makes it similar to some "good" ones in Table 6. On the other hand, the latter are among those providing good recognition accuracy. However, by comparing those samples with the problematic one, we can further notice that in Fig. 3 the high ambient illumination led to a sharp mirroring of the mobile device on user's eye (it is also possible to detect the user's hand and other objects). This condition causes the detection of a significant amount of fake iris features, increasing the entropy of the iris image and, consequently, the ambiguity of the subject's identity. Table 7 reports the dissimilarity scores when the iris on the left of Fig. 3 is compared with another one extracted from a sample belonging to the same subject. The considered comparison is with the sample on the right of the same figure, with results in Table 7. Such results point out the mentioned behaviour. They also show that, in this specific case, the recognition methods by Barra et al. and by Santos et al. achieve a good result when compared to verification scores of the other two methods, depending on the segmentation method adopted. However, the fact that some recognition methods work better than others happens quite at random, and there is no guarantee that the same methods always give a similar good verification score. It is worth reminding that the iris in Fig. 3 has been selected as an example of many similar mobile acquisitions in MICHE-I that present this kind of issues. With a deeper analysis of Table 7, it is pos-

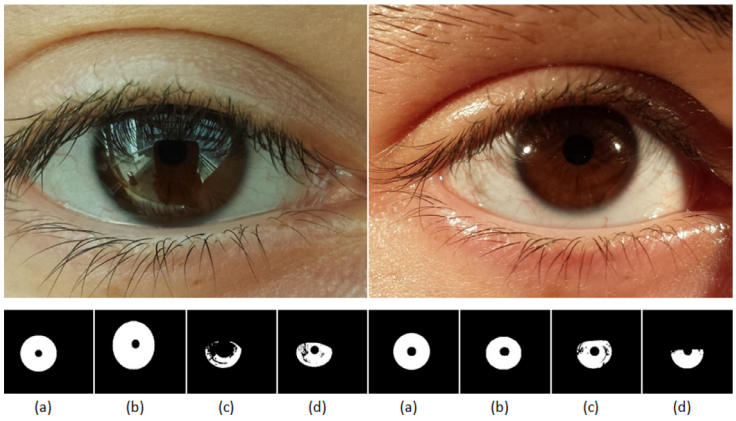

Fig. 3. Example of high resolution samples of the same subject, with good contrast and illumination in outdoor and indoor conditions, top left and top right respectively. On the leftmost sample, the high level of details produces a significant amount of fake iris features that introduce an unpredictable and unquantifiable bias in the performances of all the recognition methods. The bottom line shows the iris segmentations provided by the four methods considered in this study: (a) Barra et al. (b) Abate et al. (c) Haindl et al. (d) Yang et al.

sible to notice that the segmentation by Haindl et al. produces the most stable results when adopted by the recognition methods considered in this study. Although the corresponding verification scores are not the optimal ones, the result is interesting because it further confirms the stability/reliability of this segmentation method in comparison to the others (see Section 4.2). To this regard, let us observe that the iris mask in Fig. 3(c), left part, ignores a significant part of the region that is interested by occlusions due to environmental reflections, and, consequently, a lower amount of fake iris features is included. This justifies the stable behaviour of all four recognition methods when the segmentation by Haindl et al. is used. On the other hand, the removal of occlusions may also cause the removal of true iris patches, and therefore also reduces the amount of true iris features, thus possibly impacting on the level of verification performances. Notwithstanding this, Table 7 shows that the verification scores produced when the segmentation by Haindl et al. is used are, on average, the highest ones. This confirms that noise reduction always produces a recognition improvement by avoiding the introduction of ambiguous iris features.

Table 8
Mean processing times, including preliminary segmentation, for each method in both verification mode and identification mode.

Segment. Algorithm	Recogn. Algorithm	Verific. (1:1)	Identific. (1:150)
IS_IS	Barra	0.046 s	6.419 s
	Raja	$<$ 0.001 s	16.355 s
	Abate	0.080 s	12.889 s
	Santos	0.057 s	7.829 s
Bird	Barra	0.049 s	6.340 s
	Raja	0.001 s	33.728 s
	Abate	0.103 s	15.528 s
	Santos	0.054 s	6.705 s
Haind et al.	Barra	0.049 s	7.378 s
	Raja	0.001 s	26.297 s
	Abate	0.098 s	14.714 s
	Santos	0.067 s	10.654 s
Yang et al.	Barra	0.048 s	7.320 s
	Raja	0.001 s	18.980 s
	Abate	0.070 s	12.460 s
	Santos	0.050 s	6.990 s

5.2. Execution time

A recognition operation carried out directly on a mobile device (without the intervention of a remote server) would be mostly limited to verification cases, with template(s) of a single subject stored locally. Though having sufficient computational resources, which nowadays is still not realistic for large scale applications, privacy and security issues would suggest avoiding to maintain on a mobile device an entire template gallery. Therefore, in case of identification (1:N matching without any identity assumption), segmentation and feature extraction could still be carried out locally, to avoid transmitting the full biometric sample, so preserving user's privacy and integrity of the biometric trait. However, after this preliminary step, the acquired probe template would be transmitted to a remote recognizer, running either on a desktop or a server architecture. It is interesting to estimate the execution time for an identification operation limited to this case, and to assess if such execution time is linearly related with a single verification operation, as expected, i.e., if methods are really scalable. This section reports the average identification time to compare a probe image versus 150 gallery images. The latter seems a reasonable number to estimate the relation between the time required by a single operation and by a set of similar ones. According to this, it is possible to measure the level of performance of each segmentation+recognition method explored in this study. Though randomly picked, the images used in this experiment grant a wide variety of conditions (indoor, outdoor, frontal or rear camera, eyeglasses, makeup, shadows and so on) thus obtaining an overall mean rate of performance for each method. All images belong to the set captured by iPhone5 (the device providing images with a median resolution among those used). Table 8 reports the time values for each segmentation method evaluated. All tests were carried out on an Intel Xeon X5482 CPU 3,20GH (dual core) 64bit, 10GB RAM. The pre-processed images resolution is of 400×300 pixels for the recognition methods by Abate et al. and Santos et al. (which process the whole image with the corresponding mask). The normalized irises have 512×64 pixels for methods by Barra et al. and Raja et al., working on the normalized iris only.

All recognition methods present a quite stable response time in the verification mode, notwithstanding the adopted segmentation. However, identification mode causes some interesting differences in response times that, in this mode, seem to significantly depend on the segmentation. For instance, the method by Raja et al. requires a more than double time for identification passing from IS_IS to BIRD segmentation. While time differences fade for a single match operation, they become more significant with larger scale comparisons. It is further possible to observe that the method by Raja et al. is definitely the one providing the shortest response time in verification, notwithstanding the adopted seg-

mentation method, while it unpredictably becomes the slowest one in identification, where the fastest one is, on the average, the method by Barra et al.

6. Score level fusion

This section provides the results obtained by score level fusion of the recognition results observed for the different methods. Each experimental session was identified by the pair of capture devices involved, namely either the same device for probe and test or two different devices, by the segmentation method, and by the recognition method(s) exploited (either a single one or a score level fusion of the results from a possible subset). Each session produced a distance matrix that, for each pair of images probe/gallery, contains the corresponding score in terms of dissimilarity (the lower, the higher the probability that the two irises are from the same subject). We experimented a multi-expert approach by fusing the results from all recognition methods considered in this study, or from subsets of them. This is expected to improve the overall accuracy of the final system, since flaws in one method can be compensated by strengths in another one. In order to perform a proper fusion of the results at score level , the values returned by the different methods had to be normalized so to fall within in the same numerical range, typically [0, 1], and be comparable. The *min-max* normalization rule was used. Let *DM* be a distance matrix $m \times n$ containing the dissimilarity scores s between all possible pairs of the m irises in the probe-set and the n irises in the gallery-set, and s a score, the normalized score sn is given by

$$sn = \frac{s - \min(DM)}{\max(D) - \min(D)} \quad \forall s \in DM. \tag{3}$$

Two score level fusion strategies were exploited in this study: the *Simple Sum* fusion and the*Matcher Weighting Fusion*. The former consists in just summing up the scores produced by each method. Let $s_{i,\,j,\,m}$ be the score generated by the recognition method m for the pair of images $< i, j >$, the simple sum fused score $ss_{i,\,j}$ is:

$$ss_{i,j} = \sum_{m=1}^{M} s_{i,j,m} \quad \forall i, j \tag{4}$$

where M represents the number of recognition methods whose results have to be fused. Considering that the distance matrix obtained by the sum of the single scores might be defined in a new range of values, namely [0, M], a MinMax normalization step is further carried out on the fused distance matrix in order to work with a common range of values in [0,1].

The *Matcher Weighting Fusion* makes use of the Equal Error Rate (EER) achieved by the recognition methods, and assigns a higher weight to those methods that achieve a lower EER. The weights are therefore inversely proportional to the corresponding errors of the methods considered. Let m be the recognition method and e_m its error, the weight w_m is calculated as:

$$w_m = \frac{\frac{1}{\sum_{m=1}^{M} \frac{1}{e_m}}}{e_m} \tag{5}$$

where $0 \leq w_m \leq 1$ and $\Sigma_{m=1}^{M} w_m = 1$. Once the weights have been computed, the matcher weighting fused score $f_{i,\,j}$ becomes:

$$f_{i,j} = \sum_{m=1}^{M} w_m s_{i,j,m} \quad \forall i, j \tag{6}$$

In this study, four segmentation methods were considered and, for each of them, the effectiveness of four recognition methods was compared. In each of these combinations, nine different experimental sessions have been designed either intra- or inter-device (GS4 vs GS4, GS4 vs IP5 etc), which produced a big amount of experimental results even just considering a single recognition method. The score level fusion further significantly increases the number of results to consider, due to the implementation of all possible fusion schemes, i.e., to the need to consider all the possible subsets of recognition results. In fact, given a pair of devices and a segmentation method, it is possible to exploit any subset of the four different feature extraction/matching methods, i.e., it is possible to fuse at score level the results of any out of the 11 possible subsets of the matching methods. In order to provide a comprehensible view and discussion of the achievements from data fusion, only the most relevant results are presented.(Table 9)

The following tables present the results for individual segmentation methods. Each table reports the triple (DEC, AUC, EER) for each of the nine possible combinations of devices, and for the best two fusion schemes, using either Simple Sum or Matcher Weighting respectively. By an overall view of fusion results, the improvement achieved by using any of the two fusion strategies is rather limited. In many cases the AUCs, which is not possible to report here for sake of space, are just a little wider than the ones obtained by an execution of Santos et al. algorithm alone. It is possible to appreciate the negligible difference in performance also by looking at values in Tables 4 and 5, that report the results achieved in the corresponding settings by the single recognition algorithms.

The first observation that comes out is that the increased computing power required to fuse the output of the algorithms is not counterbalanced by a significant improvement in the recognition accuracy, and in this sense it is counter-productive. This result also confirms that the methods analyzed show a very different and uncontrolled behaviour, and that in general a single method achieving very high and stable performances can alone outperform any combination of less accurate methods.(Table 10)

Abate+Santos is the fusion scheme that occurs more frequently as the best one, confirming the level of performance achieved by each one of them. (Tables 11 and 12)

7. Conclusion

Reliable biometrics on handheld devices has been gaining increasing relevance and represents an extremely challenging application for computer vision systems, due to the wildness of the environments and of the unconstrained data acquisition protocols. This paper has discussed in detail the results of MICHE-I contest, the first international contest specifically devoted to iris/ocular recognition using data acquired from multiple types of handheld devices. The paper has started by briefly summarizing the MICHE-I benchmark and the participating algorithms. Afterwards, their effectiveness and the linear correlation of their results have been compared, in order to appreciate the possible improvements due to fusion techniques.

It is worth underlining that the performance levels reported in this paper should not be compared to those achieved by other solutions for iris recognition, as the average quality of the data being used for the MICHE-I contest is far lower. Instead, the main idea in MICHE-I was to assess the feasibility of ocular recognition solutions to work in mobile settings, and to provide the first baseline results, which could be the basis for further improvements in subsequent initiatives. The MICHE-I data acquisition protocol was designed to contain images from indoor/outdoor environments, taken using the frontal/rear cameras of various devices and without particular supervision. The resulting data provided the opportunity to answer the question: *is it feasible to recognize human irises from present average mobile devices with a sufficient level of accuracy?*.

As concluding remarks, we note that particular efforts should be paidto the segmentation/quality assessment phases of the processing chain, as these phases could reduce data's heterogeneity. Also, the use of semantic information (as it is done in periocu-

Table 9
Fusion results of the IS_IS segmentation algorithm.

		GS4				IP5				GT2			
		FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER
GS4	SS												20.158 ± 0.010
		Aba+San	27.887 ± 0.272 (27.013–28.225)	20.880 ± 0.006 (20.870–20.893)	20.213 ± 0.012 (20.190–20.236)	Aba+San	26.389 ± 0.094 (26.236–26.597)	20.813 ± 0.007 (20.800–20.828)	20.264 ± 0.012 (20.240–20.285)	Aba+San	25.384 ± 0.351 (25.120–26.535)	20.906 ± 0.008 (20.896–20.926)	(20.140–20.176)
	MW	Aba+San	27.423 ± 0.242 (26.699–27.751)	20.879 ± 0.006 (20.869–20.892)	20.213 ± 0.012 (20.187–20.234)	Aba+San	26.289 ± 0.096 (26.134–26.514)	20.815 ± 0.007 (20.803–20.830)	20.265 ± 0.012 (20.239–20.287)	Aba+San	25.299 ± 0.341 (25.089–26.408)	20.905 ± 0.008 (20.893–20.922)	20.165 ± 0.007 (20.152–20.178)
IP5	SS												
		Aba+San	210.292 ± 1.022 (26.776–210.824)	20.820 ± 0.008 (20.807–20.838)	20.254 ± 0.008 (20.238–20.266)	Aba+San	210.403 ± 1.058 (27.652–211.102)	20.864 ± 0.007 (20.854–20.878)	20.221 ± 0.009 (20.198–20.236)	Aba+San	27.578 ± 0.791 (25.276–29.196)	20.839 ± 0.013 (20.820–20.870)	20.216 ± 0.012 (20.189–20.235)
	MW	Aba+San	210.834 ± 0.428 (29.617–211.576)	20.821 ± 0.009 (20.807–20.840)	20.257 ± 0.010 (20.233–20.270)	Aba+San	211.164 ± 0.563 (29.481–212.082)	20.864 ± 0.007 (20.854–20.880)	20.219 ± 0.010 (20.200–20.236)	Aba+San	27.922 ± 0.471 (27.706–29.421)	20.836 ± 0.013 (20.816–20.866)	20.215 ± 0.012 (20.190–20.233)
GT2	SS		26.130 ± 0.127				26.396 ± 0.117						20.154 ± 0.015
		Aba+San	(25.951–26.388)	20.833 ± 0.009 (20.819–20.853)	20.226 ± 0.012 (20.202–20.247)	Aba+San	(26.221–26.639)	20.811 ± 0.010 (20.794–20.832)	20.241 ± 0.009 (20.220–20.258)	Raj+San	24.866 ± 1.097 (24.399–28.093)	20.919 ± 0.008 (20.908–20.936)	(20.125–20.178)
	MW	Aba+San	26.009 ± 0.126 (25.821–26.275)	20.837 ± 0.009 (20.820–20.827)	20.223 ± 0.009 (20.197–20.212)	Bar+San	25.470 ± 0.220 (25.269–26.117)	20.813 ± 0.009 (20.797–20.831)	20.252 ± 0.012 (20.231–20.276)	Raj+San	25.386 ± 0.797 (24.999–27.726)	20.919 ± 0.009 (20.908–20.942)	20.154 ± 0.016 (20.113–20.175)

Table 10
Fusion results of the BIRD segmentation algorithm.

		GS4				IP5				GT2			
		FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER
GS4	SS						26.107 ± 0.122						
		Aba+San	27.077 ± 0.206 (26.870−27.717)	20.874 ± 0.006 (20.864−20.889)	20.205 ± 0.006 (20.187−20.216)	Aba+San	(25.941−26.378)	20.836 ± 0.007 (20.826−20.850)	20.221 ± 0.008 (20.207−20.238)	Aba+San	26.279 ± 0.237 (26.031−27.042)	20.833 ± 0.008 (20.818−20.850)	20.239 ± 0.009 (20.217−20.253)
	MW	Aba+San	26.656 ± 0.216 (26.437−27.277)	20.876 ± 0.006 (20.865−20.889)	20.208 ± 0.008 (20.187−20.220)	Aba+San	25.849 ± 0.131 (25.678−26.141)	20.843 ± 0.007 (20.831−20.857)	20.213 ± 0.008 (20.199−20.229)	Aba+San	26.109 ± 0.253 (25.897−26.911)	20.832 ± 0.008 (20.819−20.847)	20.238 ± 0.008 (20.221−20.252)
IP5	SS						27.309 ± 0.189						
		Aba+San	26.244 ± 0.171 (26.052−26.734)	20.807 ± 0.008 (20.793−20.823)	20.264 ± 0.009 (20.246−20.281)	Aba+San	(26.775−27.560)	20.832 ± 0.007 (20.819−20.847)	20.224 ± 0.008 (20.207−20.238)	Aba+San	210.110 ± 1.347 (26.352−210.919)	20.796 ± 0.008 (20.782−20.812)	20.257 ± 0.010 (20.237−20.278)
	MW	Aba+San	26.071 ± 0.195 (25.860−26.628)	20.810 ± 0.009 (20.796−20.828)	20.264 ± 0.008 (20.246−20.277)	Aba+San	26.762 ± 0.189 (26.240−27.037)	20.836 ± 0.007 (20.825−20.852)	20.226 ± 0.008 (20.208−20.239)	Aba+San	210.954 ± 0.503 (29.100−211.352)	20.802 ± 0.008 (20.789−20.819)	20.262 ± 0.009 (20.245−20.277)
GT2	SS		26.708 ± 0.114										
		Aba+San	(26.532−26.973)	20.839 ± 0.008 (20.826−20.855)	20.221 ± 0.009 (20.203−20.241)	Raj+San	26.738 ± 0.142 (26.532−27.057)	20.836 ± 0.007 (20.823−20.848)	20.227 ± 0.006 (20.214−20.238)	Aba+San	26.959 ± 0.109 (26.777−27.194)	20.893 ± 0.007 (20.883−20.907)	20.178 ± 0.008 (20.159−20.190)
	MW	Aba+San	26.448 ± 0.119 (26.259−26.717)	20.842 ± 0.008 (20.829−20.861)	20.229 ± 0.011 (20.205−20.246)	Raj+San	26.797 ± 0.102 (26.623−27.008)	20.844 ± 0.007 (20.830−20.856)	20.226 ± 0.007 (20.212−20.239)	Aba+San	26.643 ± 0.109 (26.455−26.886)	20.898 ± 0.007 (20.888−20.912)	20.174 ± 0.007 (20.158−20.186)

Table 11
Fusion results of the segmentation algorithm by Haindl et al..

		GS4				IP5				GT2			
		FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER
GS4	SS		27.388 ± 0.143				27.277 ± 0.158				27.313 ± 0.225		
		Aba+San	(26.966–27.631)	20.873 ± 0.006	20.195 ± 0.008	Aba+San	(26.792–27.525)	20.834 ± 0.007	20.240 ± 0.006	Aba+San	(26.611–27.595)	20.830 ± 0.006	20.235 ± 0.007
				(20.864–20.886)	(20.181–20.209)			(20.823–20.849)	(20.228–20.250)			(20.818–20.843)	(20.221–20.250)
	MW		27.078 ± 0.155				27.114 ± 0.184				27.139 ± 0.265		
		Aba+San	(26.661–27.331)	20.877 ± 0.005	20.194 ± 0.009	Aba+San	(26.565–27.405)	20.838 ± 0.007	20.237 ± 0.010	Aba+San	(26.415–27.504)	20.837 ± 0.007	20.231 ± 0.008
				(20.869–20.889)	(20.177–20.208)			(20.828–20.852)	(20.219–20.257)			(20.825–20.851)	(20.214–20.245)
IP5	SS												
		Aba+San	25.961 ± 0.088	20.809 ± 0.007	20.264 ± 0.009	Bar+San	25.426 ± 0.056	20.856 ± 0.006	20.217 ± 0.008	Aba+San	26.828 ± 0.160	20.802 ± 0.007	20.269 ± 0.006
			(25.815–26.113)	(20.797–20.822)	(20.249–20.283)		(25.327–25.550)	(20.847–20.869)	(20.202–20.232)		(26.365–27.026)	(20.790–20.816)	(20.256–20.280)
	MW	Bar+San	25.551 ± 0.068	20.826 ± 0.006	20.257 ± 0.005	Bar+San	25.715 ± 0.069	20.864 ± 0.006	20.203 ± 0.008	Raj+Aba+San	26.455 ± 0.173	20.809 ± 0.007	20.271 ± 0.009
			(25.433–25.684)	(20.816–20.839)	(20.247–20.268)		(25.598–25.875)	(20.855–20.878)	(20.190–20.220)		(26.206–26.942)	(20.797–20.822)	(20.251–20.284)
GT2	SS						27.109 ± 0.273						
		Raj+San	26.967 ± 0.215	20.826 ± 0.008	20.250 ± 0.007	Raj+San	(26.525–27.697)	20.811 ± 0.008	20.249 ± 0.008	Raj+San	210.320 ± 0.329	20.894 ± 0.005	20.176 ± 0.005
			(26.528–27.473)	(20.813–20.841)	(20.236–20.263)			(20.797–20.828)	(20.234–20.264)		(29.993–211.073)	(20.886–20.903)	(20.165–20.185)
	MW		27.176 ± 0.139				27.588 ± 0.178						
		Aba+San	(26.969–27.475)	20.836 ± 0.006	20.245 ± 0.006	Raj+San	(27.341–28.001)	20.816 ± 0.008	20.246 ± 0.009	Raj+San	29.499 ± 0.187	20.901 ± 0.005	20.158 ± 0.004
				(20.826–20.848)	(20.233–20.257)			(20.802–20.832)	(20.226–20.265)		(29.240–29.945)	(20.893–20.911)	(20.150–20.168)

Table 12
Fusion results of the segmentation algorithm by Yang et al..

		GS4				IP5				GT2			
		FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER	FUSION	DEC	AUC	EER
GS4	SS		27.219 ± 0.198								27.129 ± 0.145		
		Aba+San	(27.023−27.805)	20.890 ± 0.005 (20.881−20.901)	20.194 ± 0.009 (20.172−20.207)	Aba+San	26.464 ± 0.139 (26.276−26.853)	20.878 ± 0.005 (20.870−20.888)	20.196 ± 0.006 (20.185−20.207)	Aba+San	(26.753−27.375)	20.857 ± 0.006 (20.847−20.870)	20.213 ± 0.007 (20.198−20.227)
	MW	Raj+San	27.782 ± 0.299 (27.545−28.798)	20.901 ± 0.005 (20.892−20.912)	20.167 ± 0.006 (20.155−20.179)	Aba+San	26.349 ± 0.123 (26.180−26.667)	20.886 ± 0.005 (20.877−20.896)	20.185 ± 0.006 (20.175−20.196)	Raj+Aba+San	26.915 ± 0.116 (26.722−27.177)	20.870 ± 0.005 (20.860−20.879)	20.215 ± 0.010 (20.194−20.232)
IP5	SS		26.519 ± 0.192										
		Aba+San	(26.218−27.028)	20.859 ± 0.006 (20.849−20.871)	20.216 ± 0.005 (20.206−20.225)	Aba+San	26.862 ± 0.088 (26.726−27.065)	20.884 ± 0.005 (20.876−20.895)	20.187 ± 0.008 (20.173−20.201)	Aba+San	26.646 ± 0.142 (26.487−27.058)	20.826 ± 0.007 (20.814−20.840)	20.246 ± 0.005 (20.237−20.256)
	MW	Raj+San	26.398 ± 0.157 (26.232−26.782)	20.866 ± 0.006 (20.857−20.882)	20.215 ± 0.007 (20.202−20.229)	Aba+San	26.447 ± 0.089 (26.301−26.632)	20.888 ± 0.005 (20.880−20.900)	20.195 ± 0.005 (20.183−20.204)	Aba+San	26.523 ± 0.131 (26.368−26.906)	20.833 ± 0.007 (20.822−20.848)	20.233 ± 0.008 (20.219−20.248)
GT2	SS						27.520 ± 0.199						
		Aba+San	27.396 ± 0.278 (26.575−27.677)	20.846 ± 0.005 (20.837−20.857)	20.228 ± 0.006 (20.216−20.238)	Aba+San	(26.975−27.812)	20.845 ± 0.005 (20.836−20.856)	20.244 ± 0.006 (20.232−20.255)	Aba+San	27.668 ± 0.287 (26.802−27.939)	20.910 ± 0.004 (20.904−20.918)	20.152 ± 0.005 (20.144−20.161)
	MW	Raj+San	27.266 ± 0.238 (26.562−27.527)	20.858 ± 0.005 (20.849−20.867)	20.225 ± 0.005 (20.213−20.235)	Aba+San	27.291 ± 0.205 (26.688−27.584)	20.853 ± 0.005 (20.844−20.864)	20.228 ± 0.006 (20.216−20.241)	Raj+San	28.685 ± 0.300 (27.703−28.980)	20.920 ± 0.003 (20.915−20.928)	20.145 ± 0.004 (20.137−20.152)

lar recognition) could play a role in further improvements in this technology. Finally, another obvious requirement will be the collection of massive amounts of labeled data from mobile devices. These would enable to implement/evaluate data-driven recognition strategies in this field, such as the presently extremely popular deep learning recognition approach.

Acknowledgements

The fourth author acknowledges the support given by FCT project UID/EEA/50008/2013.

References

[1] A. F. Abate, M. Frucci, C. Galdi, D. Riccio, BIRD: Watershed based IRis detection for mobile devices, Pattern Recognit. Lett. 57(51201) 43–51.
[2] S. Barra, A. Casanova, F. Narducci, S. Ricciardi, Ubiquitous iris recognition by means of mobile devices, Pattern Recognit. Lett. 57(73201) 66–73.
[3] A. Abate, M. Nappi, F. Narducci, S. Ricciardi, Fast iris recognition on smartphone by means of spatial histograms, in: In International Workshop on Biometric Authentication, Springer International Publishing, 2014, pp. 66–74.
[4] V. Bruni, D. Vitulano, A robust perception based method for iris tracking, Pattern Recognit. Lett. 57(80201) 74–80.
[5] J. Daugman, High confidence visual recognition of persons by a test of statistical independence, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1993) 1148–1161.
[6] M. De Marsico, M. Nappi, D. Riccio, IS_IS: Iris segmentation for identification systems,In Proceedings 2010 International Conference on Pattern Recognition – ICPR 2010 (2860201) 2857–2860.
[7] M. De Marsico, M. Nappi, H. Proença, Guest editorial introduction to the special executable issue on "mobile iris CHallenge evaluation part i (MICHE i)", Pattern Recognit. Lett. 57(3201) 1–3.
[8] M. De Marsico, M. Nappi, D. Riccio, H. Wechsler, Mobile iris challenge evaluation (MICHE)-i, biometric iris dataset and protocols, Pattern Recognit. Lett. 57(23201) 17–23.
[9] D. Gragnaniello, C. Sansone, L. Verdoliva, Iris liveness detection for mobile devices based on local descriptors, Pattern Recognit. Lett. 57(87201) 81–87.
[10] M. Haindl, M. Krupicka, Unsupervised detection of non-iris occlusions, Pattern Recognit. Lett. 57(65201) 60–65.
[11] K.P. Hollingsworth, S.S. Darnell, P.E. Miller, D.L. Woodard, K.W. Bowyer, P.J. Flynn, Human and machine performance on periocular biometrics under near-infrared light and visible light, IEEE Trans. Inf. Forensics Secur. 7(2) (601201) 588–601.
[12] M.S. Hosseini, B.N. Araabi, H. Soltanian-Zadeh, Pigment melanin: pattern for iris recognition, IEEE Trans. Instrum. Meas. 59 (4) (2010) 792–804.
[13] Y. Hu, K. Sirlantzis, G. Howells, Improving colour iris segmentation using a model selection technique, Pattern Recognit. Lett. 57(32201) 24–32.
[14] D. Martin, C. Fowlkes, D. Tal, J. J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, 2, 2001, pp. 416–423.
[15] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.
[16] P.J. Phillips, K.W. Bowyer, P.J. Flynn, Comments on the CASIA version 1.0 iris data set, in IEEE Trans. Pattern Anal. Mach. Intell. 29(10) (1870200) 1869–1870.
[17] P.J. Phillips, W.T. Scruggs, A.J. O'Toole, P.J. Flynn, K.W. Bowyer, C.L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale experimental results, IEEE Trans. Pattern Anal. Mach. Intell 32 (5) (2010) 831–846.
[18] W.K. Pratt, Digital image processing, Wiley-Interscience, New York, 1978.
[19] H. Proenca, S. Filipe, R. Santos, J. Oliveira, 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. Mach. Intell. 32 (8) (2010) 1529–1535.
[20] H. Proenca, L.A. Alexandre, The NICE.i: noisy iris challenge evaluation – part I, in: Proc. IEEE First Int–l Conf. Biometrics: Theory, Applications, and Systems (BTAS), 2007, pp. 27–29.
[21] H. Proenca, L.A. Alexandre, Introduction to the special issue on the recognition of visible wavelength iris images captured at-a-distance and on the move, Pattern Recognit. Lett. 33 (2012) 963–964.
[22] K.B. Raja, R. Raghavendra, V.K. Vemuri, C. Busch, Smartphone based visible iris recognition using deep sparse filtering, Pattern Recognit. Lett. 57 (2015) 33–42.
[23] J.B.T.M. Roerdink, A. Meijster, The watershed transform: definitions, algorithms and parallelisation strategies, Fundam. Inform. 41 (1–2) (2000) 187–228.
[24] G. Santos, E. Grancho, M.V. Bernardo, P.T. Fiadeiro, Fusing iris and periocular information for cross-sensor recognition, Pattern Recognit. Lett. 57(59201) 52–59.
[25] K. Nguyen, C. Fookes, R. Jillela, S. Sridharan, A. Ross, Long range iris recognition: a survey, Pattern Recognit. 72 (2017) 123–143. ISSN 0031-3203
[26] J.a. Neves, F. Narducci, S. Barra, H. Proena, Biometric recognition in surveillance scenarios: a survey, Artif. Intell. Rev. 46 (4) (2016) 515–541.
[27] D.M. Rankin, B.W. Scotney, P.J. Morrow, B.K. Pierscionek, Iris recognition failure over time: the effects of texture, Pattern Recognit. 45 (1) (2012) 145–150. ISSN 0031-3203
[28] S.A. Sahmoud, I.S. Abuhaiba, Efficient iris segmentation method in unconstrained environments, Pattern Recognit. 46 (12) (2013) 3174–3185. ISSN 0031-3203
[29] T.H.N. Le, M. Savvides, A novel shape constrained feature-based active contour model for lips/mouth segmentation in the wild, Pattern Recognit. 54 (2016) 23–33. ISSN 0031-3203
[30] M. Frucci, M. Nappi, D. Riccio, G.S.d. Baja, WIRE: Watershed based iris recognition, Pattern Recognit. 52 (2016) 148–159. ISSN 0031-3203