MICHE competitions: a realistic experience with uncontrolled eye region acquisition.

Silvio Barra and Maria De Marsico and Hugo Proença and Michele Nappi

1 Introduction

Any user in the world that has to access a protected service or location, or that simply wants to protect its owned devices, has to struggle with assuring a secure access to them. This is a first aspect that characterizes self-handled authentication strategies. Actually, the use of special signs, objects or passphrases goes back to the very origins of human communities. Watchwords asked by sentinels, or the 5-pointed pentagon tattooed on the palm of members of the Pythagorean school are examples of a kind of authentication often seen in the literature. The first attempt to use computer support for authentication is represented by passwords, that first appeared at the Massachusetts Institute of Technology in the mid-1960s, where a massive compatible time-sharing computer (CTSS) was used to pioneer many of the milestones of computing, including password-based authentication. In those times, a single password was sufficient to access one's virtual space and files, which afterall were the only resources to protect. Afterwards and beyond any forecasting, computers massively entered every-day life, with Internet allowing the creation of an increasing number of remote services of various kinds. This has caused both the corresponding increase of the number of passwords to use, and also the ~~grow~~ of the password theft risk, due to the increasing value of the protected resources. More and more complex and non-trivial passwords must be used. However, the more they are difficult to crack, the more they are difficult to remember. The possible alternative

Silvio Barra
University of Cagliari - Italy, e-mail: silvio.barra@unica.it

Maria De Marsico
Sapienza University of Rome - Italy, e-mail: demarsico@di.uniroma1.it

Hugo Proença
Universidade da Beira Interior - Portugal, e-mail: hugomcp@di.ubi.pt

Michele Nappi
University of Salerno - Italy, e-mail: mnappi@unisa.it

or addition represented by possession of physical objects (e.g., keys and cards) does not solve these problems. Rather, the need to keep the physical object always available when needed, and the possibility that it can be lost or stolen, may make things even worse for the users. In this awkward scenario, biometric authentication, though not being invincible, seems to provide a more "natural" alternative. The users can just exploit what they are or the way they behave to be recognized ~~an~~ granted privileges.

The other core element in present authentication scenarios is that mobile equipment is ubiquitous nowadays. Smartphones substituted old cellular phones, that in turn had replaced traditional landlines. The possibility to communicate almost wherever and whenever represents a characterizing aspect of the still ongoing technological revolution. However, the whatever dimension, that allows the new communication devices and protocols, is even more disrupting. The uses of present smart mobile devices include storing/transferring in real time almost any kind of multimedia information. Such data is often personal, and often sensitive. The exchange of sensitive information requires a twofold approach to address increasing security needs: it is both necessary to reliably identify the owner before the use of the device, and to reliably identify the user of a remote service at the moment the device connects to it. Biometrics can both enforce and make authentication simpler in conventional controlled environments. The next step is to move biometrics in uncontrolled settings, where there is no operator to guide ~~in~~ the capture of a "good quality" sample, and on mobile devices. Mobile biometric recognition is the new frontier for secure use of data and services.

It is interesting to remind the basic principles and issues that characterize biometric authentication. The paper by Clarke published in 1994 [Clarke, 1994] is among the earliest ones devoting specific attention to biometric recognition. Available means to achieve formal identification of individuals are classified as: 1) ways to merely distinguish among individuals - Names and Codes; 2) ways to verify individual identity - Knowledge-Based Identification and Token-Based Identification; and, finally, 3) biometrics, that can be used for both verification and identification. In this classification, ~~The~~ term "biometrics" refers to identification techniques relying on some physical and difficult-to-alienate characteristic. Of course, they require suitable measurements and matching strategies. Clarke further sketches a first taxonomy of biometric traits [Clarke, 1994]: 1) those based on appearance, that include the usual elements reported in any identity document, such as height, weight, color of skin, hair and eyes, visible markings, gender; race, facial hair, glasses, that are supported by photographs; 2) those based on (social) behavior, including body-signals, voice characteristics, speech style, visible handicaps, that are supported by video (or audio) recordings; 3) those based on bio-dynamics, including the way of signing and keystroke dynamics, that require specific capture strategies; 4) those based on natural physiography, including skull measures, teeth and skeletal injuries, fingerprint sets and handprints, retinal scans, vein patterns, hand geometry, and DNA; 5) those based on imposed physical characteristics, including collars, bracelets, microchips, and transponders. The paper by Jain et al. [Jain et al., 1997] simplifies this classification into two broad classes, namely physical or behavioral

traits, that are still used at present as reference. The paper further elaborates on Clarkes human identifiers to list the properties of a biometric trait. They are the well-known universality, uniqueness, collectability, performance, and circumvention. In [Jain et al., 2004] biometric traits are further classified as either supporting unique identification (hard traits, e.g., face or fingerprints), or providing information lacking sufficient distinctiveness and/or permanence to differentiate any two individuals (soft traits, e.g. demographic traits and most behavioral traits).

Notwithstanding the optimistic premises, the kind of interaction required from the biometric recognition systems may cause troubles to non-expert users, especially in unattended scenarios where no operator is there to assist during the task. In general, authentication systems are often difficult to use. Quoting from a paper published in 2001 by Sasse et al. [Sasse et al., 2001]: "The security research community has recently recognised that user behaviour plays a part in many security failures, and it has become common to refer to users as the 'weakest link in the security chain'. We argue that simply blaming users will not lead to more effective security systems." In 2000 Nielsen [Nielsen, 2000] assumes that "in the future, security will improve through biological [biometric] verification mechanisms, such as fingerprint recognition or retina scanning "; yet also alerts that "it will take time for this infrastructure to be built (and fingerprint systems won't work for some people)". The conclusion in [Sasse et al., 2001] is even more skeptical: "biometric systems may be a good fit for some user-tasks-context configurations, but not all of them." Concerns raised in 2004 [Patrick, 2004] and related to the acquisition step, are unfortunately still valid, as researchers dealing with biometric recognition know very well. ~~Finger-~~print readers may suffer from dirt, bad framing, different pressure and motion; face recognition systems are affected by PIE (pose, illumination, expression) distortions, and also by aging of the subject. Iris scanners may suffer from the bas alignment of the eye with the camera lens. These problems become dramatically critical when dealing with mobile biometrics. In this case, more problems rise because of the unattended acquisition, since the user may not be able to capture a good sample, and further be unaware of what good sample means in the different cases. The 2007 work by Sasse [Sasse, 2007] proposes an apparently obvious solution: Biometric systems should have user-friendly, intuitive interfaces that guide users in presenting necessary traits." However reliable use of selfie-biometrics is still an open problem. Mobile biometric recognition is continuously increasing its popularity, thanks to the possibility ~~to exploit~~ personal and/or wearable devices equipped with more and more accurate sensors. Mobile equipment is ubiquitous nowadays and allows capturing biometric traits anytime in any place, by incorporating all necessary hardware equipment and software applications for capturing and processing biometric data. However the capture phase still poses crucial problems. This dichotomy Figure 1 inspired MICHE (Mobile Iris CHallenge Evaluation) project.

The chapter develops as follows. Section 2 summarizes the main concepts related to iris recognition and how MICHE challenges are positioned with respect to the past and present research scenarios. Section 3 briefly describes the challenge setup with its two separate phases, and the dataset used as benchmark for evaluating participating approaches. Section 4 deals with the first MICHE-I challenge, focused

on iris segmentation. Section 5 presents the results of the following challenge, focused on iris recognition.

2 Iris recognition and MICHE challenges.

The iris is the circular structure in the central part of the eye that determines what is popularly defined as the *eye color*, but from a functional point of view it has a muscular nature that is responsible for controlling the diameter and size of the pupil (the inner black disk, actually a hole) and therefore the amount of light reaching the retina. From an optical point of view and comparing the eye to a camera, the pupil represents the aperture, while the iris has the role of the diaphragm. From the biometric recognition point of view and of the involved processing steps, it is worth reminding which are the most relevant external visible structures that contribute to characterize a human iris. The pupillary zone is the most internal part of the iris, whose edges mark the pupil boundary. These edges are well visible in light color eyes, while it may be difficult to distinguish them in very dark eyes. The latter is one of the problems to be addressed during iris region segmentation in

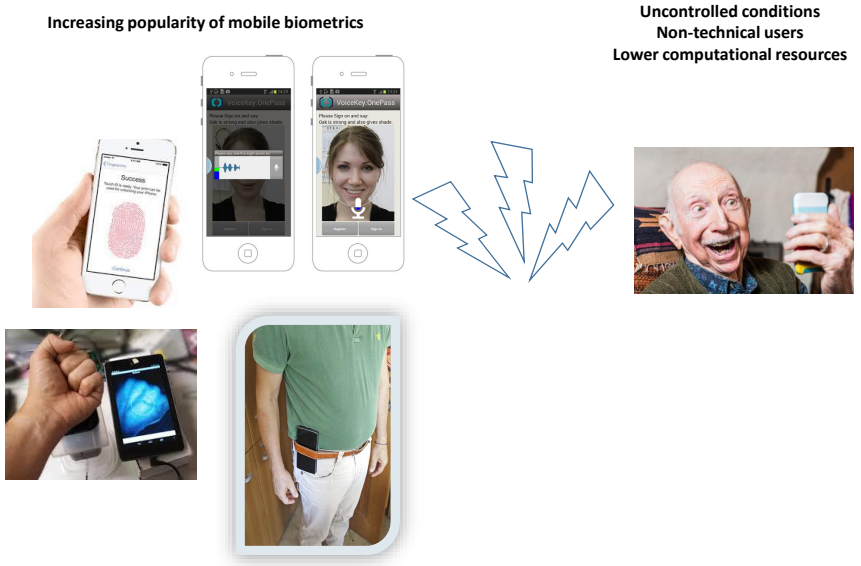

Fig. 1 Increasing popularity of mobile biometrics vs. increasing use by non-technical users.

visible light. Proceeding towards the external iris border, the collarette is a very thick region that separates the pupillary region fron the ciliary zone. In this region, the sphincter muscle and dilator muscle regulate the pupil dilation. It is relatively easy to identify this region in eyes which are not too dark, since it is made up of radial ridges extending from the periphery to the pupillary zone. The ciliary zone extends up to meet the sclera. The overall iris structure is characterized by both regularities, represented by radial furrows, as well as singularities, represented by crypts and possible lighter/darker spots (Figure 2).

The iris is among the best candidates for biometric recognition. It is extremely discriminative: right and left iris of the same person are so different to hinder a correct matching. This is due to the fact that randotypic elements largely overcome genotypic ones in individual development. In other words, contrarily to, e.g., face, the genetic baggage has very little influence on the iris makeup process. Its small size makes related image processing quite fast with respect to face, and its peculiarities make it very difficult to spoof an iris template. The most used kind of iris codes, devised by Daugman [Daugman, 1993], is among the less expensive templates from the storage point of view, and the acquisition is little intrusive. For all these reasons, the iris is a natural candidate for mobile biometric recognition.

Research results regarding related techniques have quickly progressed from the pioneering work by Daugman [Daugman, 1993] and Wildes [Wildes, 1997], mostly pertaining controlled settings and ~~Nera~~-Infrared (NIR) capture settings, to the use of deep learning [Liu et al., 2016], with the most recent NICE (Noisy Iris Challenge Evaluation) addressing iris recognition in less controlled settings [Proença and Alexandre, 2007, Proenca and Alexandre, 2012]. Mobile setting and especially the inherent problems related to uncontrolled acquisition are addressed in the two challenges of the MICHE project [De Marsico et al., 2017, De Marsico et al., 2018] whose results are the core topic of this chapter. Whatever the context, the iris recognition workflow follows the same processing steps, which are typical of any object

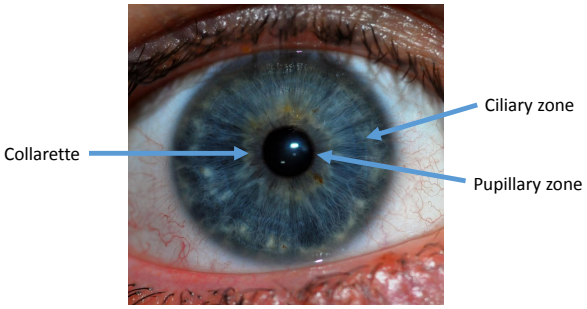

Image adapted from the original image by JDrewes - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3117810

Fig. 2 Most relevant regions of iris images, for biometric recognition purposes. The collarete divides the pupillary and cilliary zones, an is particularly visible in light-pigmented irises.

detection/recognition procedure. The ease of localizing the eyes within the faces, and the characteristic annular shape of the iris, should facilitate a reliable and accurate detection of this anatomical element and the creation of a suitable representation. This especially holds when NIR capturing is used, since reflections and illumination variations have little influence on images, and a controlled acquisition guarantees a correct position with respect to the camera. On the contrary, when capture is carried out in Visible Light (VL), while images usually contain precious chromatic features than NIR images, they are also much more seriously affected by many noisy artifacts produced by light sources and reflections, and their processing suffers from dark pigmentation. Therefore, the first difficulties soon arise when attempting to detect and segment the iris. Of course, a poor segmentation compromises all the following steps, since feature extraction would be possibly carried out on non-iris regions, while the complete set of (possibly unconnected) iris regions would not be correctly identified. Iris segmentation was the focus of the first MICHE challenge, aimed at assessing the accuracy of the candidate algorithms. It is worth noticing that segmentation not only identifies the useful iris region, but also usually produces a segmentation mask to be used during matching to leave out non-iris patches. The step following segmentation is iris sample normalization. In most approaches, this does not only entail reducing iris images to a common size, but also computing a polar representation facilitating the following processing. The most used technique to obtain this is the Rubber Sheet Model introduced by Daugman [Daugman, 2009]. The same procedure is applied to the segmentation mask Afterwards, different approaches extract and match different features, either related to the regular patterns that can be identified, or to possible singularities, either local or global [Bowyer and Burge, 2016]. Feature extraction and recognition are the focus of the second MICHE challenge. Figure 3 shows the typical steps in iris processing and recognition and points out those addressed by MICHE. Of course the kind of the final result depends on the entailed recognition modality, either verification (1:1 matching) or identification (1:N matching).

3 Challenge setup and MICHE dataset

As anticipated above, the Noisy Iris Challenge Evaluation (NICE I) addressed the problem of matching images captured in unconstrained conditions. The iris dataset used as benchmark, namely UBIRIS.v2 [Proenca et al., 2010] has been captured in the visible wavelength (VW), at-a-distance (4 to 8m), and on the move. The results confirm how VW and uncontrolled conditions together dramatically affect recognition performance. Similar conclusions result from the following NICE II contest [Proenca and Alexandre, 2012]. Notwithstanding this, MICHE project addresses a further problem. While UBIRIS datasets were acquired by high resolution cameras, MICHE dataset, as we will detail better in the following, only uses built-in cameras of different smartphones, that at the moment of capture produced images of undoubtedly lower quality than UBIRIS ones. MICHE challenges followed the same

NICE schema: a first one focusing on iris segmentation and using as benchmark MICHE dataset, and a second one focused on feature extraction and matching, carried out with an extended version of the dataset and using as a common segmentation tool the best method resulted form the first challenge.

Why was a new dataset required? The Chinese Academy of Sciences was a pioneer in collecting the first publicly available iris datasets dealing with iris images, continuously updated from CASIA-IrisV1 to CASIA-IrisV4 since 2002 [Ma et al., 2003, Sun et al., 2014]. Its images are either collected under NIR or are synthesized. Therefore, until NIR sensors will spread on mobile devices, these datasets cannot be used to assess iris processing on mobiles. Similar considerations hold for benchmarks used for ICE competition mentioned before. On the contrary, UBIRIS datasets, available from SOCIA Lab at University of Beira Interior (Portugal), though being captured in visible light and uncontrolled conditions. have a much better resolution than average mobile sensors. Figure 4 shows some details of the last CASIA versions, while Figure 5 shows a sample from a ICE competition and a sample from UBIRIS, both from a left eye. Looking at the two samples it is easy to understand the difference between the two addressed contexts.

The aim of MICHE was to assess the real feasibility of iris recognition when images are captured in visible light, by "normal" user level mobile devices, by "normal" (non necessarily technical) users in uncontrolled/unattended conditions, and when cross-device matching can be needed (Figure 6)

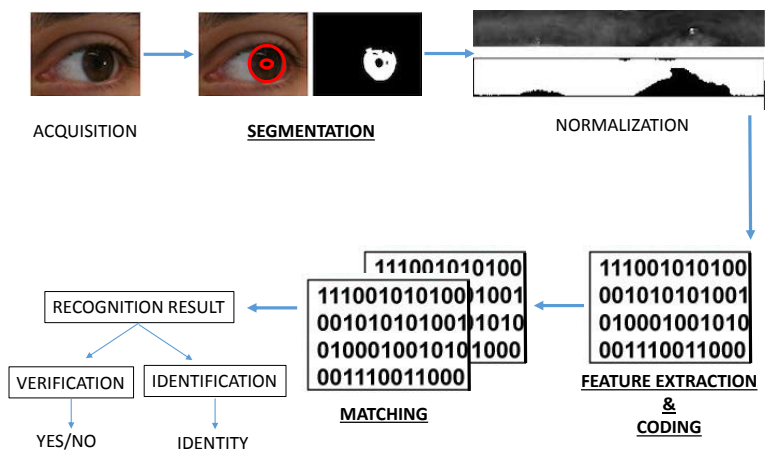

Fig. 3 The main phases of iris recognition processing chain (bold font underlines the phases addressed by MICHE challenges).

MICHE challenges provided as a benchmark a dataset reflecting this specific set up. Before continuing, it is worth pointing out two symmetrical considerations. The accuracy of capture/ quality of the captured image may be enhanced due to the usually short distance (not more than the length of a normal human arm) and to the user natural attitude to have a frontal pose while taking a selfie. In addition, in this case we are considering collaborative users, that have all interest in being recognized. The reverse of the medal is that the quality of the captured image can suffer from possible lower resolution of the mobile device camera, from motion blur and illumination distortions, from incorrect image framing, that can be all caused by either/both the kind of device, the possible lack of technical experience of the user, and by the lack of control on user capture operation. Addressing these problems requires more robust detection/segmentation and matching procedures. It is worth pointing out again that the performance of the matching can be dramatically affected by the quality of the segmentation. This is the reason for having all participants to the second part of the challenge to all use the same segmentation: they have a common starting point so that it is possible to evaluate the addition of feature extraction/matching in a fair way.

The composition of the dataset used for MICHE reflects the use of different mobile devices for the acquisition and a realistic simulation of the acquisition process including different sources of distortion/noise. The data was captured across several acquisition sessions separated in time, to get a realistic amount of intra-class variations. All images were annotated with metadata useful to carry out demographics- as well as well as device-based analysis. In order to reproduce a real-world setting, the

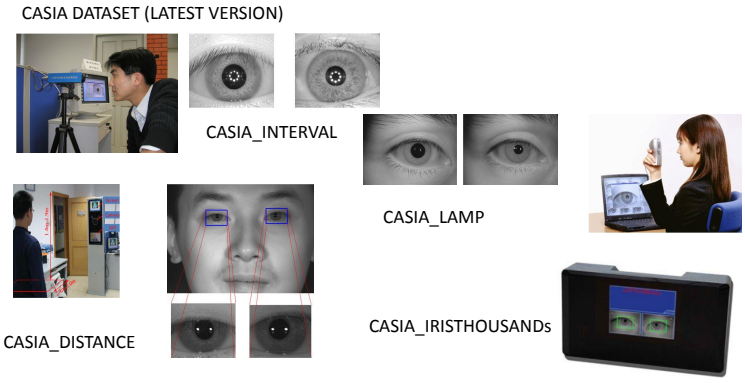

Fig. 4 Last versions of CASIA datasets.

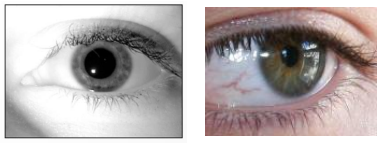

Fig. 5 A sample from ICE (left) and one from UBIRIS (right).

subjects involved in experiments were given no special instruction but were rather advised to take a selfie of their eye as they would do if asked in a real situation. For instance, subjects usually wearing eyeglasses could either remove or keep them. The self-images of their iris were acquired by normally holding the mobile device. For each session, a minimum of 4 shots for each camera (a device could possibly have two) and acquisition mode (either indoor or outdoor) was requested. Indoor acquisition was affected by various sources of artificial light, sometimes combined with natural light sources. Outdoor acquisition exploited natural light only. Only one iris per subject was acquired. The three kinds of devices used for data acquisition purposes (both smartphones and tablets) that (at the time!) were representative of the current top market category, can represent at present medium-level devices, and are listed by increasing camera resolution:

- Galaxy Tablet II (GT2)

Fig. 6 Cohesive perspective of the MICHE operational context.

  - Operating System: Google Android
  - Posterior Camera: N/A
  - Anterior Camera: 0.3 Megapixels

- iPhone5 (IP5)
  - Operating System: Apple iOS
  - Posterior Camera: iSight with 8 Megapixels (72 dpi)
  - Anterior Camera: FaceTime HD Camera with 1.2 Megapixels (72 dpi)

- Galaxy Samsung IV (GS4)
  - Operating System: Google Android
  - Posterior Camera: CMOS with 13 Megapixel (72 dpi)
  - Anterior Camera: CMOS with 2 Megapixel (72 dpi)

It is interesting to point out that it is possible to identify three groups of images at three different resolutions (1536 x 2048 for iPhone5, 2322 x 4128 for Galaxy S4, and 640 x 480 for the tablet). Examples are shown in Figure 7. This is a further challenge for cross-device matching.

Sources of noise affecting the MICHE dataset images include all those that can be present in real-world unattended settings. Different kinds of reflexes are among the most frequent ones, and can be caused by either artificial light or natural light sources, as well as by people or objects in the scene. Out of focus and blur can be either due to an incorrect capture operation or to involuntary movements of the hand and/or of the head and/or of the eye during selfie capturing. Part of the region of interest may be occluded by eyelids, eyeglasses, eyelashes, hair, or shadows. The device itself may introduce artifacts due to low resolution or sensor defects, or it may present different color dominants. Further problems are raised by off-axis gaze and variable illumination. Actually, Figure 8 shows that these factors can also affect images in in UBIRIS.v2 (http://nice2.di.ubi.pt/). However comparison of Figure Figure 7 and Figure Figure 8 is useful to further point out the features of MICHE images. Most of all, due to lack of precise framing and to different capture distances, it is possible to obtain either well centered eyes or half faces or partial eye images. Of course, this causes a different position, size and sharpness of the region of interest, i.e. the region useful for iris recognition. This happens because capture can happen from a very close distance up to a distance equal to the arm length, even if in general the users maintain an average capture distance that is in the middle. This means that more robust eye localization technques are required as a first processing step, but at the same time it is possible to use, when present, useful information from the periocular region.

The dataset is annotated by metadata through a reference XML file for each iris image. The information recorded regards image acquisition (e.g., device characteristics, distance from the device, outdoor/indoor indication) and archiving (e.g., file name and file type), subject demographics, and the conditions under which the image was acquired. At present, MICHE dataset contains images from 75 different subjects, acquired in two sessions separated in time (1 to 9 months apart) with 1297

Fig. 7 Examples of MICHE images. From top to bottom, images were taken by iPhone5, by Galaxy S4, and Galaxy Tablet II devices.

images by GS4, 1262 images from IP5, and 632 images from GT2. The dataset also contains a MICHE Fake and a MICHE Video subsets, to test Presentation Attack Detection (PAD) approaches and recognition from dynamic data. More details on the dataset can be found in [De Marsico et al., 2015]. Table 1 summarizes the difference in terms of possible distortions between controlled and uncontrolled acquisition. It is worth pointing out that ”controlled” in this context means assisted by an expert operator that can appreciate the possible defects in the obtained image

Table 1 Summary of the data degradation factors affecting image acquisition in controlled/uncontrolled conditions

	Motion blur	**Out of focus**	**Reflections**	**Occlusions (eyelids/ eyelashes)**	**Off-axis**	**Closed/ partially closed eye**	**Partial eye region (bad framing)**	**Extended periocular region (bad framing)**	**Low Resolution**
Controlled	NO	NO	YES	YES	NO	NO	NO	NO	YES/ NO
Uncontrolled	YES	YES	YES	YES	YES	YES	YES	YES	YES/ NO

and ask to repeat the acquisition until a satisfactory visual quality is achieved. Low resolution can be considered as a possibly common problem for the two settings, depending on the acquisition device. However, in uncontrolled conditions it can add to possible other distortions, and can be more frequent in mobile capture.

4 MICHE-I challenge: iris segmentation

The participants to MICHE-I challenge had the above described dataset as a common benchmark. The aim of the challenge and of the analyses carried out on its results was to explore both image covariates that are likely to cause a decrease in the performance levels of the compared algorithms, and the further effect of cross-device operations. Segmentation was the only focus of the challenge, however participants had the possibility to also integrate their proposal with a recognition module. This allowed to test both the original approaches, and possible re-combinations of segmentation (S) and recognition (R) modules.

The analysis of results went beyond the evaluation/identification of the best approaches, but also focused on the image features/distortions that can mostly positively/negatively affect the final recognition performance, and on interoperability issues caused by the use of different devices in enrollment vs. testing phases. Having

Fig. 8 Examples of UBIRIS images.

more recognition modules available, it was also possible to investigate multi-classier strategies, to complement the strengths of more different approaches. The fusion was carried out at score level either by the popular Simple Sum, or by a weighted sum policy assigning higher weights to methods achieving a lower Equal Error Rate (EER) in a pre-testing step. Returning to the challenge setting, the performance measures chosen to evaluate the different iris/non iris segmentation strategies were a set of quite classical ones used 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). The corresponding performance measures were Decidability index, Equal Error Rate (EER), and Receiver Operator Characteristic (ROC) curves with corresponding Area Under Curve (AUC). Since MICHE-I was especially focused on iris segmentation, more metrics were used to measure performance in this operation. Moreover, when present, the results of proposals also addressing iris recognition were analyzed concentrating on the segmentation methods allowing a more reliable feature extraction and matching thanks to a better separation of the eye regions.

4.1 Metrics used to evaluate the segmentation quality

Table summarizes the performance measures used to evaluate the candidate methods in MICHE-I, and that are quite common for classification problems (in our case, iris/non-iris) or even multi-class problems. Some of them are specifically suited for segmentation: Pratt metric is introduced in [Pratt, 2007] and Global Consistency Error in [Martin et al., 2001].

Each of the metrics exploited to evaluate the segmentation quality is able to capture some specific aspect of a correct classification. Of course, it is firstly important to correctly classify an existing edge pixel (true positives vs. false negatives) and this ability is different from avoiding false positives (vs. true negatives). Actually the two can be in contrast, so that the lower the rate of false negatives, the higher the number of false positives could be. This common for binary classifiers, and as a matter of fact they do not play an asymmetrical role in evaluating segmentation algorithms: an algorithm that achieves the former might be less effective to achieve the latter. Errors in either direction can differently affect the test of the processing. A contour interruption caused by false negatives can hinder if not completely compromise the detection of a shape, or produce an unconnected contour where a connected one is needed/expected. The role of the first 4 metrics is to measure these aspects separately. F1-score rather provides an overall estimate of the ability of the algorithm to distinguish true edge pixels from false ones without missing too many of them. RI is a measure of the overall agreement between positive/negative classifications and ground truth, taking into account pairs of corresponding pixels. It can be extended to more different candidate classifications. E1-Score represents a kind of complementary measure, since it rather measures the proportion of disagreeing

Table 2 The performance measures used to evaluate the methods submitted to MICHE-I

Accuracy	Accuracy measures the proportion of true results, summing up true positives and true negatives and computing the rate with respect to the total number of samples.
Precision	Precision measures the proportion of the true positives against all the returned positive results, that include both true and false positives.
Recall	Recall is also called the true positive rate, or the sensitivity, and measures the proportion of positives that are correctly identified as such, i.e. true positive against ground truth positives.
Specificity	Specificity is the true negative rate, and measures the proportion of negatives that are correctly identified as such, i.e. true negatives against ground truth negatives.
F1-Score	F1-Score can be interpreted as a weighted average of the precision p and the recall r, and is defined as: $F1-Score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$; it ranges from a best value of 1 and a worst worst value of 0
Rand Index (RI)	The Rand Index counts the fraction of pairs of corresponding pixels in given segmentation and ground truth, whose elements are either both labeled as edge or both labeled as non-edge, both in ground truth and in the returned segmentation.
E1-Score	E1-Score represents the classification error rate of the algorithm on the input image, and is given by the proportion of corresponding disagreeing pixels (that have a different label in the returned segmentation and in ground truth); it can be computed by the logical EXCLUSIVE-OR operator.
Pratt	Pratt metric is defined as a function of the distance between correct and measured edge positions; it is also indirectly related to the false positive and false negative edges: $Pratt = \frac{1}{max\{E_G,E_D\}} \times \sum_{k=1}^{E_D} \frac{1}{1+\alpha+d_i^2}$ where 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 edge point and the closest ground truth one, and α is a scaling constant that in the original metric formulation is $\alpha = \frac{1}{9}$; this metric takes into account the global trend of the distances between returned and ground truth edges; it ranges between an optimal value of 1 and a minimum of 0.
Global Consistency Error (GCE)	The Global Consistency Error evaluates at which extent one segmentation can be viewed as a refinement of the other; given two segmentations S_1 and S_2, a pixel p_i and regions $R(S_1,p_i)$ and $R(S_2,p_i)$. containing the pixel in segmentation S_1 and S_2 respectively, a local (asymmetric) error measure is defined as $E(S_1,S_2,p_i) = \frac{\lvert R(S_1,p_i) \setminus R(S_2,p_i) \rvert}{\lvert R(S_1,p_i) \rvert}$, so that it is possible to compute a local refinement error in each direction at each pixel; the Global Consistency Error forces all local refinements to be in the same direction, being finally defined as $GCE = \frac{1}{n} min\{\sum_i E(S_1,S_2,p_i), \sum_i E(S_2,S_1,p_i)\}$ with n the number of pixels; substituting the minimum of the sums with the sum of the minima provides the Local Consistency Error (LCE), that would rather allow refinement in different directions in different parts of the image.
Pearson Correlation Coefficient (PCC)	The Pearson Correlation Coefficient is a measure of the linear correlation between two random variables X and Y, returning a value between +1 and 1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

pixels. Pratt metric evaluates accuracy from a point of view more strictly related to the specific segmentation problem, since it returns a global estimate of the distance between the detected contours and the ground truth: not only true/false, not only lack of correspondence, but also distance from the true result. In this respect GCE provides a similar yet more ”directed” result, taking into account the direction of the error too: it measures how the errors with respect to ground truth (the direction is fixed) can result in a less detailed segmentation though bringing much the same core information. Finally, PCC is the usual Pearson correlation, to evaluate if the result segmentation and the ground truth present a similar trend.

4.2 Methods participating in MICHE-I

Since most proposals included both segmentation and recognition, both approaches will be briefly summarized when appropriate. A first observation deriving from the analysis of the methods is that it is a common practice to try to compensate for poor image quality using different approaches. A frequent one entails the use of the periocular region as an extra source of information. When the resolution of the iris region is not sufficient, or too many distortions are present, recognition can be supported by additional features extracted from the region around the eye. Among the other proposals along this line, this approach had been proposed also in NICE II challenge (addressing recognition), and in particular by the winning method by Tan et al. [Tan et al., 2012]. A combined approach using more sets of features/methods can reduce the specific sensitivity to any particular data covariate. Last but not least, it is possible to exploit color compensation techniques to attenuate the typical cross-device difference of sensor features. The participating methods are listed below by alphabetical order of the first author's last name.

The approach by Abate et al. [Abate et al., 2015] for the challenge relies on an algorithm based on the watershed transform for iris segmentation, namely watershed Based IRis Detection (BIRD). The first step is to compute for each RGB channel the gradients in a colored, illumination-corrected image. The final gradient image is obtained by averaging the gradients computed over the separate channels. The watershed transform exploits the topographical distance approach [Roerdink and Meijster, 2000]. The output of the watershed transform guides the binarization of the original image and the circle detection step, in order to find a parametrized expression for both the pupil and the sclera boundaries. Also this proposal exploits the the periocular region, which is localized using as reference the length of the iris radius. Differently form Santos et al, that exploit a rectangular periocular region, BIRD relies on a different choice. Starting from the approximating circle detected during the iris segmentation process, BIRD exploits its center coordinates and radius to construct two concentric ellipses that enclose part of the area around the iris. Both ellipses are centered in the center of the iris, but they are defined by different major and minor axes, always determined starting from the iris radius. The area enclosed by the ellipses is processed in a way similar to the Daug-

mans Rubber Sheet Model. The resulting rectangular region has a resolution which depends on the granularity chosen for the parameters (angle) and of the mapping. This way of processing the periocular region is quite original. For both regions feature encoding is done by means of 64-bit color histograms, matched using the cosine dissimilarity and Hamming distance. Iris and periocular results are fused at a score level, by implementing a simple sum approach.

The proposal by Barra et al. [Barra et al., 2015] also includes segmentation and recognition. The segmentation method, named IS_IS, was originally proposed in [De Marsico et al., 2010]. The original method entails using Canny edge detector to identify edges, and then the pupil boundary is identified by a voting scheme that ranks circular edges by uniformity of the inner region and contrast between the inner and outer regions. Iris is identified after applying the Rubber Sheet Model to the image, by inspecting the dark-to-light intensity variations along the columns of the image in polar coordinates. The method is modified to run on mobile devices. Feature encoding relies on spatial histograms (spatiograms). Common histograms can be considered as first order spatiograms, while higher order ones contain further information relating to the spatial domain spanned by the pixels falling in each bin. Second order spatiograms used here store also the mean and covariance matrix of the pixel coordinates. Spatiograms are matched by correlation-based techniques.

The proposal by Haindl and Krupička [Haindl and Krupička, 2015] focuses on the detection of the non-iris components, especially reflections, for the parametrizations of the iris ring. The accurate detection of eyelids and reflections can have a significant impact on the final iris segmentation. The proposed model adaptively learns its parameters on the iris texture part, and then searches for iris reflections by the recursive prediction analysis. After detecting reflections, Pupil parametrization is carried out by form fitting techniques. Next, data is converted into the polar domain according to the usual Rubber Sheet Model technique. In the resulting stripe a texture analysis phase determines the regions of the normalized data that should not belong to the iris, according to a Bayesian paradigm.

As other proposals, Hu et al. [Hu et al., 2015] apply a combined approach and fuse different iris segmentation techniques, selected according to their performance in addressing specific cases of degraded images. They proposal implements a model selection strategy, which selects the final parametrizations for iris and pupil boundaries among the candidates returned by the used baseline segmentation strategies. The selection relies on the image description provided by histograms of local gradients, that are inputted to a support vector machine providing the fused response. This strategy is designed to be modular and can be updated by adding/substituting baseline segmentation methods. This proposal does not entail either coding or classification.

The method submitted to the challenge by Santos et al. [Santos et al., 2015] entails both segmentation and recognition, and exploits all three techniques mentioned above. It uses both the information from the iris and from the periocular region, encoded/matched in a localized way. The first step is iris ring segmentation, which is carried out according to a variation of the integro-differential operator by Daugman. As a matter of fact, the characterizing part of this proposal is the encoding/matching

step. Once identified, this ring also allows locating the periocular region-of-interest (ROI). According to a combined strategy, information encoding exploits a family of texture descriptors used separately in the iris ring and in the regions surrounding the cornea (i.e., eyelids, eyelashes, skin and eyebrows). In particular, the periocular region undergoes a twofold examination, entailing both a distribution-based analysis of patches defined over a fixed grid, and a global analysis of the whole region. The former is carried out by computing both local binary patterns (LBP), histogram of oriented gradients (HOG), and Uni- form LBP (ULBP). Each descriptor is computed separately for each patch and quantized into histograms. Global analysis rather entails feature extraction from the whole periocular ROI. In this case the descriptors applied are scale-invariant feature transform (SIFT) and GIST (a set of five descriptors originally introduced in [Oliva and Torralba, 2001] to model the shape of a scene in a way that bypasses the segmentation and the processing of individual objects or regions). Iris information is encoded according to the classical approach described by Daugman [Daugman, 1993]. It may appear that much more information is captured and stored from the periocular region than from the iris. During matching, scores from all the adopted descriptors are fused by a nonlinear supervised neural network. It is worth pointing out that the method further exploits device-specific calibration techniques, that compensate for the different color rendering characterizing each experimental setup. The latter is especially useful in cross-sensor tests. In summary, this method both uses information from two different sources, the iris and the periocular region, and further uses different descriptors to fully exploit their different characteristics. The overall proposal is a good example of how a difficult setting can be addressed by an ensemble of techniques.

4.3 Some interesting notes on achieved results

We will not report the detailed results of the competition. The interested reader can refer to [De Marsico et al., 2018]. The same will be done for the results from MICHE-II. Rather, we will underline some interesting aspects as possible guidelines to take into account.

For each method and for each device, the segmentation was carried out on images captured both indoor and outdoor (OUT); in some cases no segmentation at all was returned, and tgis represents a kind of Failure To Enroll (FTE) error. The method by Haindl and Krupička achieves the highest rate of successfully segmented images, while the method used by Barra et al. achieves the lowest. However, thanks to the number of different performance measures exploited, it was possible to observe that the usable segmentation results returned by the latter, although less in number, were more accurate, providing the highest level of similarity with the ground truth. From the point of view of similarity with ground truth, the second method achieving the best results was the one by Abate et al. Therefore these two methods should provide higher quality masks, since they apply a more strict quality criterion for the obtained segmentation. Surprisingly enough, the methods by Haindl Krupička and by Hu

et al. were instead more reliable in terms of the rate of success in the following recognition step. This seems a contradiction, therefore a more careful investigation was carried out entailing the comparison of the 50 best common segmentations from the different methods. When working on the pictures where the segmentation task is easier, the results are different from those above. The mean scores on all devices testify that the method by Haindl and Krupička is actually the most reliable one. This means that it is not possible to predict the behavior of any method when problematic samples are submitted.

It is interesting to have a look to Figure 9 to appreciate the differences that can be observed in the segmentation masks, whose accuracy can be influenced by the capture condition (indoor vs. outdoor), by the resolution of the image, determined by the capture device, and by the segmentation method.

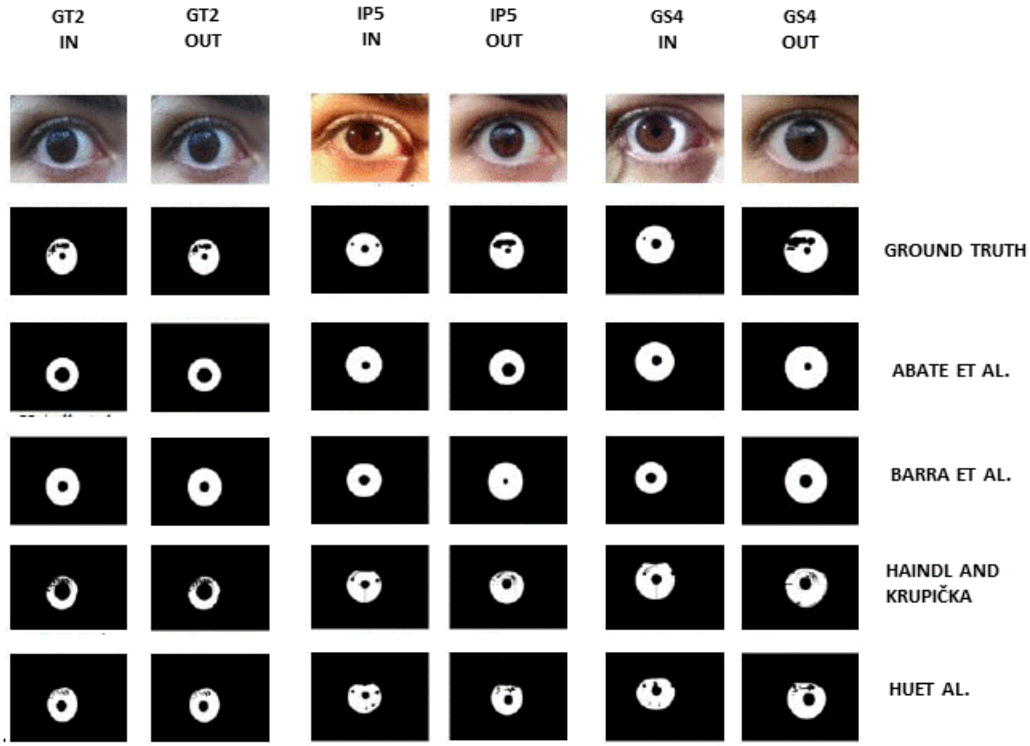

Fig. 9 Examples of the segmentation result in good quality MICHE images.

It is possible to observe the frequent degradation of image quality passing from indoor to outdoor conditions, caused by a huger presence of reflections. However, indoor capture can be influenced by the different color temperature of the illumina-

tion sources and of the sensor (see for example IP5 IN and GS4 IN in Figure 9). A higher resolution can be desirable to capture finer details. However, also the amount of noise can be greater. An example is given by the more stable segmentation produced on GT2 images. Of course, a final assessment can only be provided by the recognition results obtained on images captured in corresponding conditions.

4.4 Recombination of segmentation and recognition modules

Recognition was not specifically addressed in MICHE-I. However, since most submitted proposals also included a recognition module, this allowed to carry out further evaluations. We recombined segmentation and recognition modules in different ways. The aim was to evaluate how the two parts of a proposal depended on each other, and which segmentation methods provided a more robust preliminary step for different codings. The well-separable and not already published segmentation modules were those included in the proposals by Abate et al. [Abate et al., 2015], Barra et al. [Barra et al., 2015], Hu et al. [Hu et al., 2015], and Santos et al. [Santos et al., 2015], that have been described above.

The performances achieved by the recognition methods included in the proposals submitted to MICHE-I were evaluated in verification mode (1:1 matching, where it is to intend that a probe subject is matched against a single gallery subject, though possibly exploiting more templates per subject). The used Figures of Merit (FOMs) were decidability, area under curve (AUC) with reference to the Receiver Operating Characteristic (ROC) curve, and equal error rate (EER). The preliminary identification of the iris ROI was carried out in turn using the segmentation methods of the challenge proposals. Decidability is the same FoM used for the NICE II competition [Proença and Alexandre, 2012]. The first step to compute it requires to carry out a "one-against-all" comparison for each image $I = I_1;\ldots;I_n$ of the data set. The matching process exploits the segmented images and the corresponding binary maps $M = M_1;\ldots;M_n$ that provide the noise-free iris region identified by the segmentation step. The comparison provides a set of intra-class dissimilarity values $DI = DI_1;\ldots;DI_k$, with k the number of image pairs belonging to a same iris, and a set of inter-class dissimilarity values $DE = DE_1;\ldots;DE_m$, with m the number of image pairs belonging to different irises. The decidability value $d'(DI_1;\ldots;DI_k;DE_1;\ldots;DE_m) \rightarrow [0;\infty[$ used as evaluation measure is computed separately for each recognition method as:

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

where *avg* and σ^2 have the conventional meaning of average and variance functions computer over the parameter sets.

The challenge proposals including a clearly separable recognition module were those by Abate et al. [Abate et al., 2015], Barra et al.[Barra et al., 2015], and San-

tos et al. [Santos et al., 2015]. A further recognition method was submitted for a special issue based on but non limited to the challenge, namely the one presented by Raja et al. in [Raja et al., 2015]. It was tested in combination with all the four segmentation modules in order to get a wider set of experiments. This approach to feature extraction and recognition is based on deep sparse filtering. Sparse filtering [Ngiam et al., 2011] is an unsupervised algorithm which does not explicitly aim to model the distribution of data. It optimizes a simple cost function of sparsity using l_2 normalized features. The only parameter required in learning sparse filters is the number of features, as the sparse filters are learned by optimizing sparsity in feature distribution. Extending the method to deep sparse filters, a variable number of layers form the building blocks in learning. In the proposed approach, the deep sparse filter consists of two layers such that layer 1 is trained using 200,000 random patches of size 16 16 pixels from 4212 natural images. The sparse filtered features obtained as output from the layer 1 are normalized and provided to layer 2 using a feedforward network. The sparse filter is trained at layer 1 with 256 filters of dimension 16 16 features and at layer 2 with 256 sparse filters of 16 16 features. The sparse filter features obtained from layer 2 are exploited to extract features from the iris images. Each iris image is convolved with the 256 filters of layer 2, so that a total of 256 response images are obtained. These images are binarized, and pooled at pixel feature level in groups of eight so to obtain a response image from each pool. Afterwards an histogram is extracted from each obtained image and histograms are chained to form a feature vector.

When recognition is assessed in a context like the one addressed by MICHE, it is significant to also test cross-device performance. Segmentation can affect the final process by a different accuracy in identifying pixel regions belonging to the iris. Some such regions may be missing or some non-iris patches may erroneously enter the feature extraction and matching step. The first kind of errors become critical if quite extended, so that relevant information may be left out from the matching. This may either cause a FA (the missing region was a highly characterizing one) or a FR. The second hypothesis is less frequent,because the remaining part of the iris could be sufficient for a positive recognition. On the other hand, the second type of error may affect the final result even if less extended, since it introduces differences between two irises even where pixels should not have been considered for the matching, therefore erroneously increasing the differences. The most frequent consequence is a FRR, since in general the non-iris regions have a structure significantly different from the iris regions of both the same eye or of different eyes. Differently from segmentation, feature extraction and matching relies on finer details, that allow to summarize the microstructure of the iris region. In this case artifacts and sensor-typical noise introduced by a sensor can cause a higher accuracy degradation when captured images are matched against those captured by a different sensor. In particular, the Sensor Pattern Noise (SPN) as defined in [Lukas et al., 2006] is so specific of each device, though of the same brand and model of others, that it could be used to identify the one that captured an image.

Cross-sensor matching experiments were implemented by alternatively using sets of images acquired by the same device as either probe set (gallery) or as test set,

including intra-device recognition. Each combination of probe-gallery devices will be referred as a *class off comparison*. Detailed report of experimental results can be found in [De Marsico et al., 2018]. It is interesting here to just underline the main observed aspects.

The first observation deriving from the inspection of the results is that the recognition method by Santos et al. systematically outperforms the others in all classes of comparisons and with all segmentation algorithms. Among the classes of comparison, those entailing the same device generally allow better performance with respect to heterogeneous pairs. The class of comparison GT2vsGT2 achieves a higher level of performance (on average and compared to the others) in terms of both the EER values achieved by the various combinations of segmentation/recognition, and of the relationship between FARs and FRRs (better ROC curves). This happens notwithstanding the poorer resolution of the embedded camera. However this class generally presents the lowest decidability values. The apparent contradiction may be caused by the fact that the sizes of probe and gallery sets for GT2 are smaller than the others. This means a lower percentage of intra-classes uncertainty, which contributes to increase the level of performance, and to a lower intra-class generalizability of the results. Another non obvious observation is that performances are sensitive to the swap of the probe/gallery role of images from different sources. Of course this depends on the different level of detail of images from devices with different resolution. In general, the higher the resolution of the probe (the amount of details) with respect to the gallery images, the worse the result, because part of the probe information does not get matched. Finally, the segmentation methods by Haindl and Haindl and Krupička , and by Hu et al. provide more stable results, that cause less performance difference in the following recognition step., even if in both cases the superiority of the recognition by Santos et al. if even more evident.

A final analysis on the results of MICHE-I on when using a single recognition system was carried out to identify the ”intrinsic” covariates that can mainly affect recognition. For this reason, ”extrinsic” covariates were neglected. In particular, for each experiment the device was fixed for both probe and gallery, to neglect factors related to the device difference. Moreover also the segmentation and the recognition methods were fixed, in order to neglect the differences in the achieved similarity given by the specific techniques. The aim was to identify the best/worst pairwise comparisons that were common to all experiments. For each experiment, each gallery template is compared with each of the others of the same subject, and the full set of the obtained intra-subject dissimilarity scores is organized in a list ordered by ascending values. Given device and methods peculiarities, such scores may fall in different ranges and have different distributions across the experiments. However, it is still worth comparing the obtained rankings. The samples considered as the ”best” ones, always appear on the heading part of the ordered lists, meaning that they always achieve a very good similarity when compared with samples of the same subject. The contrary holds for ”worst” samples. The possible recurrent features of the latter are the most interesting, because they represent those intrinsic conditions that can hinder a correct recognition. Figure 10 shows some typical examples of ”worst” samples.

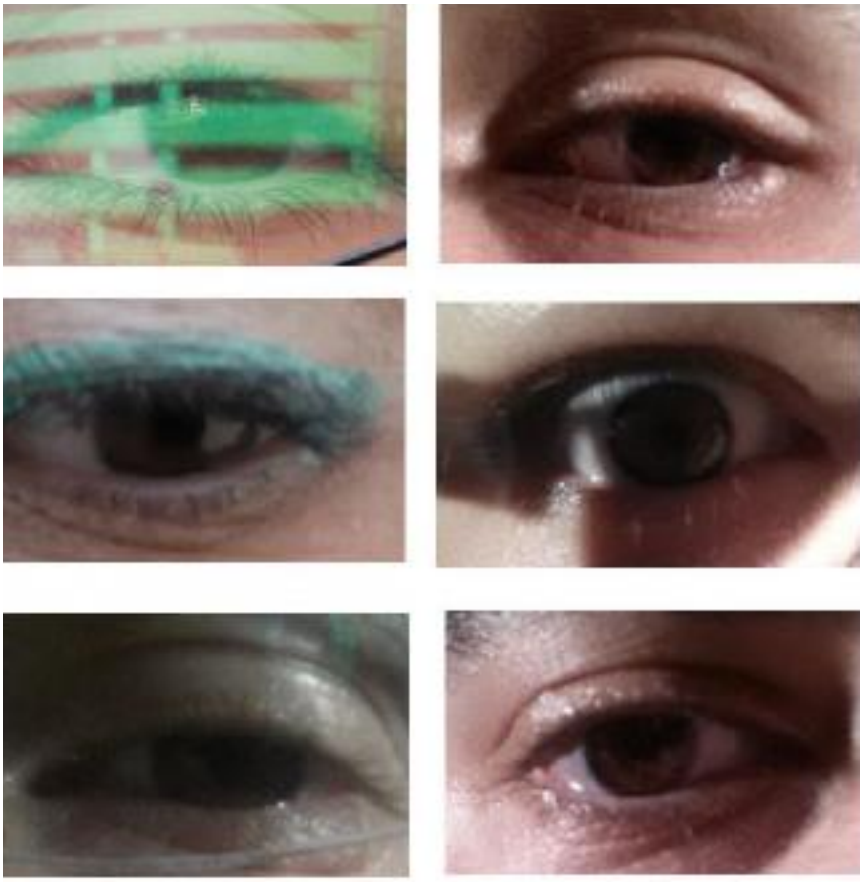

Fig. 10 Examples of ”worst” samples.

It is possible to observe that the occlusions by the eyelids are rather evident in most of the pictures; the average brightness of images is low, or the iris falls in a shadow region; reflections of unpredictable nature can affect images captured outdoors (image in the upper left corner of Figure 10. On the contrary, in ”good” samples the visibility of the irises and of the pupils is high, thus making it easier to detect and segment them.

Aiming at a possible gain in accuracy, the combination of multiple recognition methods was evaluated. In this last round of experiments, the different recognition results were fused at score level. Each experimental session was identified by the pair of (possibly different) capture devices capturing gallery and probe, by the segmentation method, and by the recognition strategy exploited. The latter can entail either a single method or a score level fusion of the results from a possible subset of them. For each session, each recognizer involved produced a dissimilarity matrix:

the lower the value of a cell (score), the higher the probability that the images on the row and on the column depicted two irises from the same subject. In order to fuse more results, the values in the matrices had to be normalized, in order to obtain comparable values in a common range [0,1]. This was achieved by the Min/Max rule, by considering the minimum and the maximum value for each matrix. Two score level fusion strategies were investigated: the *Simple Sum* fusion and the *Matcher Weighting* fusion. The former consists in just summing up the scores produced by each of the M methods involved in a session. The values in the obtained distance matrix are normalized again to remain in the range [0,1]. The *Matcher Weighting* fusion assigns to each matcher m a weight w_m that is inversely proportional to the achieved EER e_m and is defined as follows:

$$w_m = \frac{\frac{1}{\sum_{m=1}^{M} \frac{1}{e_m}}}{e_m}, \tag{2}$$

where $0 \leq w_m \leq 1$ and $\sum_{m=1}^{M} w_m = 1$.

The number of segmentation methods, of recognition methods, of pairs of probe/gallery devices, and the consequent number of their combinations, makes the amount of results to analyze and report extremely large. It is easy to guess how this amount further increases by introducing possible combinations of recognition methods in multimodal strategies, and possible different fusion strategies for each such combination. Once again, we report here only the most relevant outcomes. Given the 4 segmentation method,and for each of them the 9 combinations of probe/gallery devices, an overall analysis of the fusion results testifies that the improvement achieved by using any of the two fusion strategies is rather limited. In many cases the AUCs are just a little wider than the ones obtained by an execution of Santos et al. algorithm alone. This means that the four recognition methods taken into account have not sufficiently complementary ability to extract and match relevant features. In other words, they rely on similar information content, though represented in different ways. The increased computational demand required by running different methods and by the fusion of their results is not positively counterbalanced by a significant enough improvement in the recognition accuracy. In conclusion, a generally well performing method can achieve better performance than the combination of weaker ones, if the latter do not represent different kinds of information so to balance each other flaws.

5 MICHE-II challenge: iris recognition

Along the line of NICE challenges, the second round of MICHE challenge, namely MICHE-II, focused on iris recognition. As already underlined, the accuracy of the encoding in correctly extracting relevant and discriminative features, and the follow-

ing recognition, can be generally heavily affected by the quality of the segmentation. In order to provide a common starting point to all participants, not only a common benchmark was provided, that represents an extension of the previous dataset though maintaining the same feature distribution and variety. As for the send phase, all the competing methods had to start from the results of the same segmentation algorithm, in order to be able to assess the net contribution on the feature extraction/recognition alone. As for NICE, the best segmentation algorithm from MICHE-I was chosen, namely the one by Haindl and Krupička [Haindl and Krupička, 2015]. Of course, different feature extraction procedures can produce different templates, and specific approaches to similarity/distance evaluation. Therefore, the competitors were free to choose a suitable distance measure for the produced iris templates, with the only constraint to be a semi-metric. The higher is the dissimilarity, the higher is the probability that the two irises are from different subjects. Given I the set of images from the MICHE database, and I_a and $I_b \in I$, the dissimilarity function D had to be defined as:

$$D : I_a \times I_b \rightarrow [0;1] \subset R, \tag{3}$$

with properties

$$D(I_a;I_a) = 0 \tag{4}$$

$$D(I_a;I_b) = 0 ! I_a = Ib \tag{5}$$

$$D(I_a;I_b) = D(I_b;I_a) \tag{6}$$

Each algorithm had to return a full dissimilarity matrix among probe and gallery sets. New images were added and distance matrices were computed from scratch during the evaluation of the methods, in order to avoid any kind of bias while creating the final rank. Distance matrices were used to compute the classical FoMs to rank them, namely Recognition Rate (RR) for identification, and Receiver Operating Characteristic (ROC) curves, in particular the Area Under Curve (AUC), for verification.

5.1 Methods participating in MICHE-II

We summarize below the main characteristics of the participants. As for MICHE-I, the methods are listed by alphabetical order of the first author's last name, even if they were assigned an identifying label. Also in this case a special issue following the challenge hosted a further method that was evaluated and compared in a second round.

The methods labeled as *irisom* was implemented is first described in [Abate et al., 2017a] and experiments are extended in extended in [Abate et al., 2017c]. It implements iris recognition in the visible spectrum through unsupervised learning by means of Self

Organizing Maps (SOMs). The proposed method starts with a first step of image enhancement by simple image processing techniques, like contrast enhancement and histogram adjustment. Then it exploits unsupervised learning by Self Organizing Maps (SOMs). The SOM network clusters iris features at pixel level, after discarding those marked as non-iris in the segmentation mask. The discriminative feature map is obtained by using RGB data of the iris combined with the statistical descriptors of kurtosis and skewness, computed at pixel level in a neighborhood window of size 3×3. The network produces a feature map with the activation status of the neurons for each pixel. The map represents a cluster decomposition of the image, which maps the problem of iris recognition onto a lower dimensional space. The method then computes the Histogram of Gradients (HOG) over the obtained feature maps, and the result is used as a feature vector. Verification relies on the Pearson correlation coefficient computed in the [0,1] real interval. The best results for this method were achieved with 5×5 and 10×10 SOMs.

The method *otsedom* is described in [Aginako et al., 2016a], and experiments are extended in [Aginako et al., 2017a]. The proposal was submitted by a joint team from Universidad del Pais Vasco (UPV) and Universidad de Las Palmas de Gran Canaria (ULPGC). The approach combines popular Computer Vision techniques and Machine Learning paradigms. Starting from the segmentation, well known local descriptors are computed. Those suitable to the problem are selected after evaluating a collection of 15, with different grid configuration setups. Popular examples of the selected descriptors are Local Binary Patterns (LBP), Local Phase Quantization (LPQ), and Weber Local Descriptor (WLD). They are used individually to build separate classifiers by a supervised Machine Learning approach. Each classifier computes the dissimilarity between two irises by the histogram distance between the two a-posteriori probability distributions computed from the two iris images. In a second step, the best combination of subsets of classifiers are evaluated to build the best multi-classifier system out of the individual ones. In practice, the final algorithm combines the best five descriptors to obtain a robust dissimilarity measure of two given iris images. The mode of each a-posteriori probability for each class value is used to combine the classifiers. Some combinations of local descriptors also take into account periocular region.

A group with a slightly different composition from the above presented a further set of proposals collectively labeled as *ccpsiarb* [Aginako et al., 2016b]. The experiments presented by the authors are extended in [Aginako et al., 2017b]. Based on the training dataset given by MICHE II, a set of classifiers is constructed and tested, aiming at classifying a single image. Iris images are processed using well-known image processing algorithms. Different transformations of the original pictures can highlight different characteristics of the images. Examples of the transformations tested are Equalization, Gaussian, Median, etc. This phase aims at expressing the variability in the aspect of a picture, so to obtain different values for the same pixel (feature) positions. The output image are considered the input of the previously trained classifiers, obtaining the a posteriori probability for each of the considered

class values. The classifiers implement some well known ML supervised classification algorithms, with completely different approaches to learning: IB1, NaiveBayes, Random Forest and C4.5. Experiments take into account the 19 image collections obtained by applying single transformations, and the four different classifiers, giving a total of 76 experiments. After testing all these combinations, the Edge transformation followed by IB1 classification (identified as combination *ccpsiarb_17*) is identified as the combination providing the best results.

The *tiger_miche* method described in[Ahmed et al., 2016], with experiments extended in extended in [Ahmed et al., 2017], uses a combination of a popular iris code approach and a periocular biometric based on the Multi-Block Transitional Local Binary Patterns. To generate iris codes, the method convolves the unwrapped iris image with 1-D Log-Gabor filter. Log-Gabor functions are chosen because they have no DC component, and this can alleviate the negative influence of light intensity differences on textural information, which affects the images captured in the visible spectrum. Since a 1-D filter is used, each row in the unwrapped image is treated as 1-D signal. It is multiplied in a frequency domain with 1-D Log-Gabor filters of different scales, that capture textural information with different level of details. To generate the iris code, the phase information of the output signals is quantized into four levels, one for each possible quadrant in the complex plane. The coding then discards iris code values at positions corresponding to either very small or very large amplitudes of filter response. Hamming distance is used to match the iris codes, once adapted to take the segmentation mask into account.

Transitional LBP (TLBP) uses comparisons between neighbor pixels in a clockwise direction for all pixels, except the central one, so that it encodes encodes information about the partial ordering of border pixels. Its formulation is

$$TLBP_{P,R} = s(g_0, g_{P-1} + \sum_{i=1}^{P-1} s(g_i - g_{i-1})2^i, \tag{7}$$

where, as usual, g_0 is the central pixel of the window over which the code is computed, P is the number of neighbors, R is the window radius, ans $s(x)$ returns 1 or 0 according to the sign of its argument. Multi-block extensions of both LBP and TLBP use the average gray values from the blocks of pixels instead of the gray values of individual pixels to create the code. The method uses block sizes 3×3, 9×9, and 15×15. For each block size, it computes $TLBP_{12,3}$ and $TLBP_{24,6}$ codes and their histograms, which are concatenated to create a feature vector. Histogram vectors are matched using Chi-Square distance between the concatenated histograms.

The Hamming distance between two iris codes and the periocular matching score are computed separately, and then combined by a score-level fusion to improve the system accuracy. The values returned by the matchers fall in different ranges and present a very different score distributions, therefore the authors exploit z-score normalization.

The method labeled as *karanahujax* is described in [Ahuja et al., 2016], with experiments extended in [Ahuja et al., 2017]. It exploits a hybrid convolution-based model, for verifying a pair of periocular images containing the iris. The baseline proposed model is based on Root Scale-Invariant Feature Transform (SIFT). The binary mask is used to get the iris image rid of occlusions. Then Dense color Root SIFT descriptors [Arandjelovic and Zisserman, 2012] are computed, giving keypoints with identical size and orientation. The hybrid model is conceived as a combination of this baseline model and of two deep networks, an unsupervised one and a supervised one. The unsupervised convolution-based deep learning approach (Model1) uses a stacked convolutional architecture, with external models learned a-priori on external facial and periocular data, on top of the baseline Root SIFT model: the approach is completed by different score fusion strategies. The supervised approach (Model2) also uses a stacked convolution architecture but the feature vector is learned in a supervised manner. The fusion carried out in the hybrid model exploits an average of the computed scores after suitable normalization.

FICO_matcher exploits the FIRE (Fast Iris REcognition) algorithm is described in [Galdi and Dugelay, 2016]. Related experiments are extended in extended in [Galdi and Dugelay, 2017]. The key features of the method are the use of a combination of classifiers exploiting the iris color and texture information, and its limited computational time, that makes it particularly attractive for fast identity checking on mobile devices. The classifiers whose results are fused respectively exploit the distance among color, texture, and "cluster" features, meaning the presence of specific pixel aggregations in the image. In order to compute color distance, given two irises, each picture is first split into small blocks. For each pair of corresponding blocks, the color distance is computed, and the minimum color distance obtained is the final score re-turned by the color descriptor. The exploited color distance measures is the KolmogorovSmirnov distance. Given the cumulative histograms of images, with $\hat{h}_i = \sum_{j \leq i} h_j$, the distance is defined as:

$$d_{K-S}(H,K) = \max_i(|\hat{h}_i - \hat{k}_i|). \tag{8}$$

The texture descriptor relies on the MinkowskiBouligand dimension (box-counting dimension). The box-counting dimension of a set S is defined as:

$$dim_{box}(S) = lim_{\varepsilon \to 0} \frac{logN(\varepsilon)}{log\frac{1}{\varepsilon}}, \tag{9}$$

where where $N(\varepsilon)$ is the number of boxes of side length ε required to cover the set S. Images are decomposed in layers according to the colors, each layer is divided into blocks and for each of them the box-counting dimension is computed. These are finally chained in a feature vectors, and matching relies on Euclidean distance.

"Clusters" are connected components resulting from morphological operators applied to image layers obtained as for texture description. The features characterizing

clusters are centroid coordinates, orientation and eccentricity. Such cluster features vectors are chained to make up an image feature vector.

Two versions of the method are tested, namely V1 that uses of three kinds of descriptors, by suitably waiting the obtained distances, and V2 that does not use texture.

The method *Raja* described in [Raja et al., 2017] did not participate in the first round of the MICHE-II competition, but was submitted for the following special issue and was therefore tested from scratch together with the others. It proposes deep sparse filtering carried out on both multiple image patches and on the complete image. The image corresponding to each RGB channel is divided into a number of blocks. Both such blocks and the whole image are processed to obtain deep sparse histograms using the set of deep sparse filters. The final feature vectors is the concatenation of the set of histograms obtained from different channels and blocks. The extracted features are represented in a collaborative subspace, to jointly represent the set of training samples that correspond to enrollment. In such space a new classification approach is adopted.

5.2 Some interesting notes on achieved results

The ranking of the participant methods was obtained by running all the methods from scratch at BipLab - University of Salerno, over an extended set of images after segmenting them with the segmentation algorithm provided for the competition. The final rank list in Table 1 reports the best performing version among the ones submitted for each author (label). The rank was obtained by averaging the Recognition Rate (RR) and the Area Under Curve (AUC) achieved, and considering only images captured by the two smartphones. Both cross-device (ALLvsALL) and single-device settings were considered

- *tiger_miche*
- *karanahujax_Model2*
- *Raja*
- *irisom_10x10*
- *FICO_matcher_V1*
- *otsedom*
- *ccpsiarb_17*

As for the segmentation results, details can be found in [De Marsico et al., 2017], while it is interesting here to point out some interesting aspects of the outcomes.

As a first observation. the better the ranking achieved, the more stable the method with respect to the test setting. Of course, the hardest conditions are those found in ALLvsALL, where gallery and probe images come from different devices in unpredictable pairings. As expected, all methods provided consistently lower performances in this condition. The results confirmed the observation stemming from

MICHE-I results: the images over which the best results were achieved in homogeneous settings (gallery and probes from the same device) come from IP5, and the achieved results further present a lower standard deviation, notwithstanding the lower resolution of the camera. Once more this seems to suggest that, in the given uncontrolled and noisy conditions, higher resolution may also increase the way the noise typical of iris images can affect recognition. A related observation regards the way the different methods behave with respect to the different devices. The best method achieves high results with both cameras. Four methods, namely *karanahujax_Model2, irisom_10x10, FICO_matcher_V1* and *ccpsiarb_17* rather achieve their best performance with IP5. Methods developed in more versions are more stable w.r.t. the different variations. For instance, *karanahujax_Model1* and *karanahujax_Model2* achieve the same final score, but *Model2* achieves the better behavior in ALLvsALL. A similar constant behaviour is observed for the many versions of *ccpsiarb*, while *FICO_matcher_V2* achieves dramatically worse results than *FICO_matcher_V1*, confirming the expected outcome that texture distance is critical for iris matching. Execution times were not evaluated in the competition, but are important for real-time operations. The best method *tiger_miche* also achieved the best result in terms of time required by the single matching operation. Only *FICO_matcher_V1* did better, and even more *FICO_matcher_V2*, that, however, provides much lower recognition accuracy. On the other extreme we find the methods relying on ML techniques, which therefore seem not suited for a real time operational setting.

6 MICHE after the challenges

The previous sections have shown the role of the MICHE dataset within the challenges using it as benchmark: robust approaches have been designed, developed and tested both for segmentation and recognition/verification purposes, mainly thanks to the dual nature of the dataset itself:

- the different acquisition modalities adopted for the enrollment of the subject, from the indoor/outdoor acquisition to the different illumination conditions, and most of all the different of devices, have allowed the design and testing of cross-sensor verification algorithms;
- from a different point of view, the capture protocol assured a well-balanced presence of images presenting all the possible distortions that can affect iris images in realistic mobile unattended conditions.

In the last years, pattern recognition performance both in terms of accuracy and of computing time have been considerably improved, mainly due to the wide diffusion of the artificial intelligence-based approaches like fuzzy-controllers configurations and Machine/Deep Learning techniques. Therefore, it has become possible to address more complex problems and conditions, also in biometrics. As a consequence, despite the high level of complexity of the images in the MICHE dataset,

and thanks to its characteristics, a number of researchers investigating iris recognition have used it also outside the challenge to train and/or test their architectures over these images, also reaching quite interesting results. These approaches have successfully addressed each of the issues involved in a typical iris recognition systems, as summarized in Figure 11.

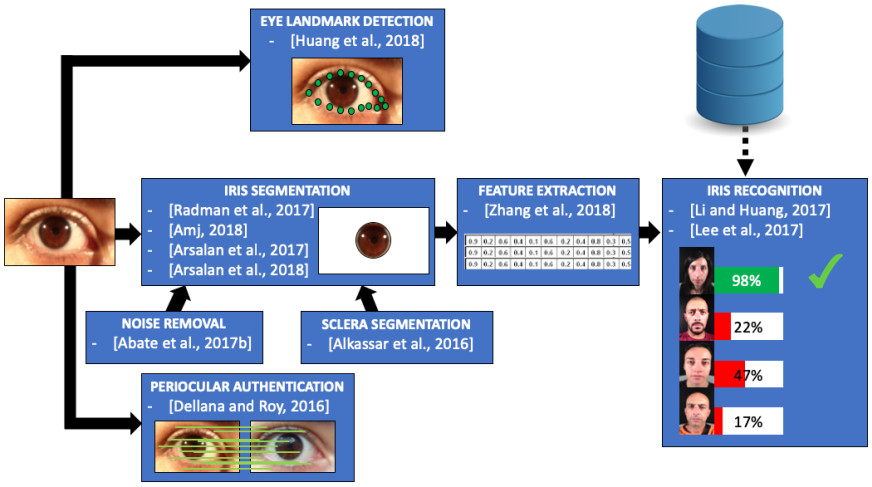

Fig. 11 A schema organizing the works that have used MICHE dataset according to the specific goal of the research.

Eye Landmark Detection The proposal in [Huang et al., 2018] deals with a novel approach for eye landmarks detection with two-level cascaded convolutional neural networks. The network at the first level utilizes eye state estimation as an auxiliary task to provide the initial positions of the eyes. The shallower network at the second level fine tunes eye positions by taking as input some small regions centered at predicted eye points locations.

Noise Removal. The goal of the work presented in [Abate et al., 2017b] is implement an effective lightweight fuzzy-based solution for noise removal from iris images, which allows a fast yet reliable segmentation approach which preserves the original resolution of the iris images.

Sclera Segmentation. Sclera segmentation can represent a preliminary step for either a correct iris identification or for further processing based on the segmented area. The paper [Alkassar et al., 2016] proposes a new sclera quality measure and a method for sclera segmentation under relaxed imaging constraints. In particular, the quality measure is based on a focus measure. The sclera segmentation is obtained by

fusing the information about pixel properties of both the sclera area and of the skin around the eye. The authors also propose a template rotation for sclera alignment, and distance scaling methods to minimize the error rates when noisy eye images are captured at-a-distance and on-the-move, together with overcoming head pose rotation.

Iris Segmentation. The method in [Radman et al., 2017] accurately localizes the iris by a model relying on the Histograms of Oriented Gradients (HOG) descriptor and on a Support Vector Machine (SVM) classifier, namely HOG-SVM. Based on the achieved localization, the iris texture is automatically extracted by means of a cellular automaton which evolves via the GrowCut technique.

The study in [Arsalan et al., 2017] proposes a two-stage iris segmentation scheme based on a convolutional neural network (CNN); which is capable of accurate iris segmentation in severely noisy environments of iris recognition by visible light camera sensor.

The same group proposes in [Arsalan et al., 2018] a densely connected fully convolutional network (IrisDenseNet), able to determine the true iris boundary even with low-quality images. The approach ensures an improved information flow between the network layers, by introducing dense connectivity, i.e., the direct connections from any layer to all subsequent layers in a dense block. The experiments are carried out on five datasets, acquired in both visible and NIR light, including MICHE.

The segmentation method proposed in [Amjed et al., 2018] is designed for the unconstrained environment of the Smartphone videos. It is based on the preliminary choice of the best frames from the videos. Then it tries to enhance the contrast of these frames between dark and light regions by applying two fuzzy logic membership functions on the negative image.

Feature Extraction. The proposal in [Zhang et al., 2018] deals with a nonlinear dynamic data analysis tool, global preserving kernel slow feature analysis (GKSFA). This tool is able to extract the high nonlinearity and inherently time-varying dynamics of batch process, but, being an unsupervised feature extraction method, it lacks the ability to utilize batch process class label information. The authors propose a novel batch process monitoring method based on the modified GKSFA, namely discriminant global preserving kernel slow feature analysis (DGKSFA), which integrates discriminant analysis and GKSFA. MICHE dataset is used to exemplify discriminant and cluster analysis, to help explaining the proposed nonlinear contribution plot.

Iris Recognition. The paper [Li and Huang, 2017] proposes an iris recognition mechanism to solve the problem of user authentication in wearable smart glasses. Given the premises, the contribution deals with both hardware and software. As for the hardware, a set of internal infrared camera modules is designed, including an infrared light source and a lens module, which is able to take clear iris images within 25 cm. As for the software, the devised iris segmentation algorithm is devised to be

used on smart glasses devices. Regarding the iris recognition, the authors propose an intelligent Hamming distance (HD) threshold adaptation method which dynamically fine-tunes the HD threshold used for verification according to empirical data collected. The research in [Lee et al., 2017] proposes a new recognition method for noisy iris and ocular images by using one iris and two periocular regions, both centered in the pupil and with a slightly different radius. The approach exploits three convolutional neural networks (CNNs).

Periocular Authentication. The experiments in [Dellana and Roy, 2016] apply a Convolutional Neural Network (CNN) to carry out periocular authentication on two datasets. Several different data augmentation techniques are tried to increase accuracy, and the results testify their relative benefits.

Miscellanea. MICHE has bee exploited as benchmark even for experimenting algorithms out of the scope of the ”hard” biometric recognition (individual subject recognition), like in [Rattani et al., 2017], in which a feasibility study of gender recognition from ocular images has been proposed. In an even wider scope, given the multiple cameras involved in the acquisition process of MICHE dataset, some works have used it to assess sensor identification methods. The purpose in that case is to classify the images according to the sensor that shot it. As an example, [Kauba et al., 2018] and [Freire-Obregon et al., 2018] propose two approaches with this goal. They are respectively based on deep learning networks and on a technique based on photo-response non-uniformity noise (PRNU). Sensor features have also been occasionally exploited for binding the identity of a subject to the information related to the sensor of his/her smartphone, as shown in Figure 12, in order to obtain a double check of the user fusing biometrics and hardware metrics. Interesting detailed analyses are reported in [Galdi et al., 2015] and [Galdi et al., 2016].

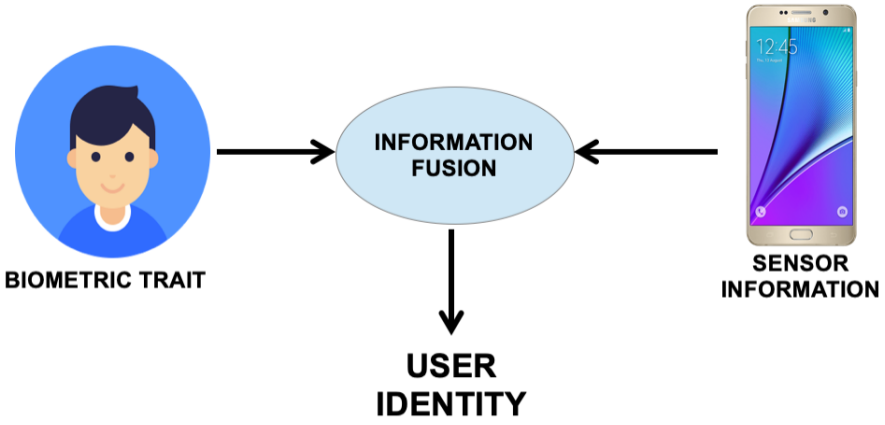

Fig. 12 The fusion of the information related to the subjects and those related to the smartphone sensor could both improve the verification of user identity and confirm the ownership of the device.

7 Conclusions

This chapter addressed the chalenges and difficulties in performing reliable biometric recognition, using self-acquired images from the subjects attempting to assess top a resource (*selfies*). In particular, we described the MICHE dataset, used as main data source for two international competitions about segmentation/recognition effectiveness of biometrics systems in such type of data. Based on our MICHE experience, it is possible to identify a number of take-home messages, presented below in the form of a list:

- An experienced operator could control specific critical conditions (e.g., pose, illumination, eye framing), possibly repeating the sample capture. However, this is not possible in uncontrolled/unattended conditions;
- The acquisition of the iris using visible light and in uncontrolled conditions presents peculiar difficulties, but it may rise even more problems when the operational setting entails a mobile application: it is necessary to compensate for users' lack of technical experience/ability, poor image quality, and also consider the possibly different features of the devices used for enrollment/recognition;
- Indoor conditions usually ~~rise~~ less illumination distortions with respect to outdoor, where a higher number of illumination sources may affect the image quality. On the other hand, data yielding from different environments has high probability of being heterogenous, due to the different color temperature of illumination sources;
- Reflections are more evident in outdoor than in indoor environments, but a more diffused and uniform illumination can create better conditions for localization and segmentation;
- Higher resolutions increase the amount of information ~~colected~~, but also increase the levels of noise. Hence, the signal-to-noise ratio appears to be weakly correlated with the resolution of the images acquired;
- It is not possible to fully and reliably predict the behavior of any method when problematic samples are submitted, i.e., there is substantial amount of work to be done in terms of reliability and robustness of recognition in case of severely degraded samples;
- The periocular region, coded either by the same or by different descriptors than those used for the iris, can improve the recognition accuracy by providing additional information. as a matter of fact, using this multi-trait strategy has become a quite used solution, especially when expecting poor quality eye samples;
- The combined use of multiple types of features can reduce the negative effect of a particular data covariate. However, at the same time, it tends to augment the computational complexity of the recognition chain, which might be particularly problematic for the execution in mobile devices;
- Typical cross-sensor differences may modify the iris micro-texture and possibly introduce artifacts, but several problems can be addressed by suitable color compensation techniques;

- It appears that the higher the resolution of the probe with respect to the gallery images, the higher the amount of unmatched information and therefore ~~the~~ lower the recognition accuracy. This observation suggests that a gallery update should be carried out when the sensor technology improves too dramatically;
- The increase of noise due to higher resolution might be limited to uncontrolled conditions where no capture adjustment is attempted, as in the case of the MICHE dataset. Due to the lack of extensive cross-resolution tests in either controlled or uncontrolled conditions, it is not possible to generalize this observation;
- Intrinsic factors affecting the recognition problem (not related to either the capture device or the segmentation/recognition methods) are the iris occlusions due to eyelids, the low brightness of the samples, the existence of shadows in the iris region, and reflections of unpredictable shape and color inside the iris ring;
- Fusion of different features and ~~different~~ classifiers can improve the matching phase, when each component highlights and takes into account a different relevant aspect for coding and matching. However:
  - fusing more recognition methods can be effective only if they take into account sufficiently complementary information; this may not be true notwithstanding the different way of representing features, if the information content is basically the same;
  - it is not sufficient to fuse different computer vision techniques to enhance the image and different descriptors to capture different properties; it is also necessary to identify those processing steps able to extract the really relevant information.
- Machine-Learning based techniques seem still too demanding, especially in terms of the computational time cost, to be exploited in real-time operations in mobile devices, where the computing power is limited and the requirement of low energy consuption is a strong constraint.

References

[Abate et al., 2017a] Abate, A., Barra, S., Gallo, L., and Narducci, F. (2017a). Skipsom: Skewness & kurtosis of iris pixels in self organizing maps for iris recognition on mobile devices. pages 155–159. Institute of Electrical and Electronics Engineers Inc.

[Abate et al., 2017b] Abate, A. F., Barra, S., Fenu, G., Nappi, M., and Narducci, F. (2017b). A lightweight mamdani fuzzy controller fornoise removal on iris images. In Battiato, S., Gallo, G., Schettini, R., and Stanco, F., editors, *Image Analysis and Processing - ICIAP 2017*, pages 93–103, Cham. Springer International Publishing.

[Abate et al., 2017c] Abate, A. F., Barra, S., Gallo, L., and Narducci, F. (2017c). Kurtosis and skewness at pixel level as input for som networks to iris recognition on mobile devices. *Pattern Recognition Letters*, 91:37 – 43. Mobile Iris CHallenge Evaluation (MICHE-II).

[Abate et al., 2015] Abate, A. F., Frucci, M., Galdi, C., and Riccio, D. (2015). Bird: Watershed based iris detection for mobile devices. *Pattern Recognition Letters*, 57:43–51.

[Aginako et al., 2017a] Aginako, N., Castrilln-Santana, M., Lorenzo-Navarro, J., Martnez-Otzeta, J. M., and Sierra, B. (2017a). Periocular and iris local descriptors for identity verification in

mobile applications. *Pattern Recognition Letters*, 91:52 – 59. Mobile Iris CHallenge Evaluation (MICHE-II).

[Aginako et al., 2017b] Aginako, N., Echegaray, G., Martnez-Otzeta, J., Rodrguez, I., Lazkano, E., and Sierra, B. (2017b). Iris matching by means of machine learning paradigms: A new approach to dissimilarity computation. *Pattern Recognition Letters*, 91:60 – 64. Mobile Iris CHallenge Evaluation (MICHE-II).

[Aginako et al., 2016a] Aginako, N., Martínez-Otzerta, J., Sierra, B., Castrillón-Santana, M., and Lorenzo-Navarro, J. (2016a). Local descriptors fusion for mobile iris verification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 165–169. IEEE.

[Aginako et al., 2016b] Aginako, N., Martínez-Otzeta, J. M., Rodriguez, I., Lazkano, E., and Sierra, B. (2016b). Machine learning approach to dissimilarity computation: Iris matching. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 170–175. IEEE.

[Ahmed et al., 2016] Ahmed, N. U., Cvetkovic, S., Siddiqi, E. H., Nikiforov, A., and Nikiforov, I. (2016). Using fusion of iris code and periocular biometric for matching visible spectrum iris images captured by smart phone cameras. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 176–180. IEEE.

[Ahmed et al., 2017] Ahmed, N. U., Cvetkovic, S., Siddiqi, E. H., Nikiforov, A., and Nikiforov, I. (2017). Combining iris and periocular biometric for matching visible spectrum eye images. *Pattern Recognition Letters*, 91:11 – 16. Mobile Iris CHallenge Evaluation (MICHE-II).

[Ahuja et al., 2016] Ahuja, K., Islam, R., Barbhuiya, F. A., and Dey, K. (2016). A preliminary study of cnns for iris and periocular verification in the visible spectrum. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 181–186.

[Ahuja et al., 2017] Ahuja, K., Islam, R., Barbhuiya, F. A., and Dey, K. (2017). Convolutional neural networks for ocular smartphone-based biometrics. *Pattern Recognition Letters*, 91:17 – 26. Mobile Iris CHallenge Evaluation (MICHE-II).

[Alkassar et al., 2016] Alkassar, S., Woo, W.-L., Dlay, S., and Chambers, J. (2016). Sclera recognition: on the quality measure and segmentation of degraded images captured under relaxed imaging conditions. *IET Biometrics*, 6(4):266–275.

[Amjed et al., 2018] Amjed, N., Khalid, F., Rahmat, R. W. O. K., and Madzin, H. B. (2018). Noncircular iris segmentation based on weighted adaptive hough transform using smartphone database. *Journal of Computational and Theoretical Nanoscience*, 15(3).

[Arandjelovic and Zisserman, 2012] Arandjelovic, R. and Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918. IEEE.

[Arsalan et al., 2017] Arsalan, M., Hong, H. G., Naqvi, R. A., Lee, M. B., Kim, M. C., Kim, D. S., Kim, C. S., and Park, K. R. (2017). Deep learning-based iris segmentation for iris recognition in visible light environment. *Symmetry*, 9(11).

[Arsalan et al., 2018] Arsalan, M., Naqvi, R. A., Kim, D. S., Nguyen, P. H., Owais, M., and Park, K. R. (2018). Irisdensenet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light camera sensors. *Sensors*, 18(5).

[Barra et al., 2015] Barra, S., Casanova, A., Narducci, F., and Ricciardi, S. (2015). Ubiquitous iris recognition by means of mobile devices. *Pattern Recognition Letters*, 57:66–73.

[Bowyer and Burge, 2016] Bowyer, K. W. and Burge, M. J. (2016). *Handbook of iris recognition*. Springer.

[Clarke, 1994] Clarke, R. (1994). Human identification in information systems: Management challenges and public policy issues. *Information Technology & People*, 7(4):6–37.

[Daugman, 2009] Daugman, J. (2009). How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier.

[Daugman, 1993] Daugman, J. G. (1993). High confidence visual recognition of persons by a test of statistical independence. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1148–1161.

[De Marsico et al., 2010] De Marsico, M., Nappi, M., and Daniel, R. (2010). Is_is: Iris segmentation for identification systems. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2857–2860. IEEE.

[De Marsico et al., 2018] De Marsico, M., Nappi, M., Narducci, F., and Proença, H. (2018). Insights into the results of miche i-mobile iris challenge evaluation. *Pattern Recognition*, 74:286–304.

[De Marsico et al., 2017] De Marsico, M., Nappi, M., and Proença, H. (2017). Results from miche ii–mobile iris challenge evaluation ii. *Pattern Recognition Letters*, 91:3–10.

[De Marsico et al., 2015] De Marsico, M., Nappi, M., Riccio, D., and Wechsler, H. (2015). Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols. *Pattern Recognition Letters*, 57:17–23.

[Dellana and Roy, 2016] Dellana, R. and Roy, K. (2016). Data augmentation in cnn-based periocular authentication. pages 141–145. Institute of Electrical and Electronics Engineers Inc.

[Freire-Obregon et al., 2018] Freire-Obregon, D., Narducci, F., Barra, S., and Castrilln-Santana, M. (2018). Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*.

[Galdi and Dugelay, 2016] Galdi, C. and Dugelay, J.-L. (2016). Fusing iris colour and texture information for fast iris recognition on mobile devices. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 160–164. IEEE.

[Galdi and Dugelay, 2017] Galdi, C. and Dugelay, J.-L. (2017). Fire: Fast iris recognition on mobile phones by combining colour and texture features. *Pattern Recognition Letters*, 91:44 – 51. Mobile Iris CHallenge Evaluation (MICHE-II).

[Galdi et al., 2015] Galdi, C., Nappi, M., and Dugelay, J.-L. (2015). Combining hardwaremetry and biometry for human authentication via smartphones. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9280:406–416.

[Galdi et al., 2016] Galdi, C., Nappi, M., and Dugelay, J.-L. (2016). Multimodal authentication on smartphones: Combining iris and sensor recognition for a double check of user identity. *Pattern Recognition Letters*, 82:144–153.

[Haindl and Krupička, 2015] Haindl, M. and Krupička, M. (2015). Unsupervised detection of non-iris occlusions. *Pattern Recognition Letters*, 57:60–65.

[Hu et al., 2015] Hu, Y., Sirlantzis, K., and Howells, G. (2015). Improving colour iris segmentation using a model selection technique. *Pattern Recognition Letters*, 57:24–32.

[Huang et al., 2018] Huang, B., Chen, R., Zhou, Q., and Yu, X. (2018). Eye landmarks detection via two-level cascaded cnns with multi-task learning. *Signal Processing: Image Communication*, 63:63 – 71.

[Jain et al., 2004] Jain, A. K., Dass, S. C., and Nandakumar, K. (2004). Soft biometric traits for personal recognition systems. In *Biometric authentication*, pages 731–738. Springer.

[Jain et al., 1997] Jain, A. K., Hong, L., Pankanti, S., and Bolle, R. (1997). An identity-authentication system using fingerprints. *Proceedings of the IEEE*, 85(9):1365–1388.

[Kauba et al., 2018] Kauba, C., Debiasi, L., and Uhl, A. (2018). Identifying the origin of iris images based on fusion of local image descriptors and prnu based techniques. volume 2018-January, pages 294–301. Institute of Electrical and Electronics Engineers Inc.

[Lee et al., 2017] Lee, M. B., Hong, H. G., and Park, K. R. (2017). Noisy ocular recognition based on three convolutional neural networks. *Sensors*, 17(12).

[Li and Huang, 2017] Li, Y.-H. and Huang, P.-J. (2017). An accurate and efficient user authentication mechanism on smart glasses based on iris recognition. *Mobile Information Systems*, 2017.

[Liu et al., 2016] Liu, N., Zhang, M., Li, H., Sun, Z., and Tan, T. (2016). Deepiris: Learning pairwise filter bank for heterogeneous iris verification. *Pattern Recognition Letters*, 82:154–161.

[Lukas et al., 2006] Lukas, J., Fridrich, J., and Goljan, M. (2006). Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214.

[Ma et al., 2003] Ma, L., Tan, T., Wang, Y., and Zhang, D. (2003). Personal identification based on iris texture analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1519–1533.

[Martin et al., 2001] Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). 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*, volume 2, pages 416–423. IEEE.

[Ngiam et al., 2011] Ngiam, J., Chen, Z., Bhaskar, S. A., Koh, P. W., and Ng, A. Y. (2011). Sparse filtering. In *Advances in neural information processing systems*, pages 1125–1133.

[Nielsen, 2000] Nielsen, J. (2000). Security and human factors. *Alertbox (November 2000)?http://www. useit. com/alertbox/20001126. html*.

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

[Patrick, 2004] Patrick, A. S. (2004). Usability and acceptability of biometric security systems. In *Financial Cryptography*, page 105.

[Pratt, 2007] Pratt, W. K. (2007). *Digital image processing*, volume 4-th edition. Wiley-interscience Hoboken, New Jersey.

[Proença and Alexandre, 2007] Proença, H. and Alexandre, L. A. (2007). The nice. i: noisy iris challenge evaluation-part i. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–4. IEEE.

[Proença and Alexandre, 2012] Proença, H. and Alexandre, L. A. (2012). Introduction to the special issue on the recognition of visible wavelength iris images captured at-a-distance and on-the-move. *Pattern Recognition Letters*, 33(8):963–964.

[Proenca and Alexandre, 2012] Proenca, H. and Alexandre, L. A. (2012). Toward covert iris biometric recognition: Experimental results from the NICE contests. *IEEE Transactions on Information Forensics and Security*, 7(2):798–808.

[Proenca et al., 2010] Proenca, H., Filipe, S., Santos, R., Oliveira, J., and Alexandre, L. A. (2010). The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535.

[Radman et al., 2017] Radman, A., Zainal, N., and Suandi, S. (2017). Automated segmentation of iris images acquired in an unconstrained environment using hog-svm and growcut. *Digital Signal Processing: A Review Journal*, 64:60–70.

[Raja et al., 2015] Raja, K. B., Raghavendra, R., Vemuri, V. K., and Busch, C. (2015). Smartphone based visible iris recognition using deep sparse filtering. *Pattern Recognition Letters*, 57:33–42.

[Raja et al., 2017] Raja, K. B., Raghavendra, R., Venkatesh, S., and Busch, C. (2017). Multi-patch deep sparse histograms for iris recognition in visible spectrum using collaborative subspace for robust verification. *Pattern Recognition Letters*, 91:27 – 36. Mobile Iris CHallenge Evaluation (MICHE-II).

[Rattani et al., 2017] Rattani, A., Reddy, N., and Derakhshani, R. (2017). Gender prediction from mobile ocular images: A feasibility study. Institute of Electrical and Electronics Engineers Inc.

[Roerdink and Meijster, 2000] Roerdink, J. B. and Meijster, A. (2000). The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta informaticae*, 41(1, 2):187–228.

[Santos et al., 2015] Santos, G., Grancho, E., Bernardo, M. V., and Fiadeiro, P. T. (2015). Fusing iris and periocular information for cross-sensor recognition. *Pattern Recognition Letters*, 57:52–59.

[Sasse, 2007] Sasse, M. A. (2007). Red-eye blink, bendy shuffle, and the yuck factor: A user experience of biometric airport systems. *IEEE Security & Privacy*, 5(3).

[Sasse et al., 2001] Sasse, M. A., Brostoff, S., and Weirich, D. (2001). Transforming the ?weakest link??a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3):122–131.

[Sun et al., 2014] Sun, Z., Wang, L., and Tan, T. (2014). Ordinal feature selection for iris and palmprint recognition. *IEEE Transactions on Image Processing*, 23(9):3922–3934.

[Tan et al., 2012] Tan, T., Zhang, X., Sun, Z., and Zhang, H. (2012). Noisy iris image matching by using multiple cues. *Pattern Recognition Letters*, 33(8):970–977.

[Wildes, 1997] Wildes, R. P. (1997). Iris recognition: an emerging biometric technology. *Proceedings of the IEEE*, 85(9):1348–1363.

[Zhang et al., 2018] Zhang, H., Tian, X., Deng, X., and Cao, Y. (2018). Batch process fault detection and identification based on discriminant global preserving kernel slow feature analysis. *ISA Transactions*, 79:108 – 126.