Multimed Tools Appl (2017) 76:23383–23411
DOI 10.1007/s11042-016-4085-8

Leveraging implicit demographic information for face recognition using a multi-expert system

Maria De Marsico[1] · Michele Nappi[2] · Daniel Riccio[3] · Harry Wechsler[4]

Received: 19 April 2016 / Revised: 18 September 2016 / Accepted: 20 October 2016 /
Published online: 18 November 2016

Abstract This paper describes a novel biometric architecture to implement unsupervised face recognition across varying demographics. The present proposal deals with ethnicity, gender and age, but the same strategy can be crafted for any mix of soft/hard biometrics, sensors, and/or methods. Our aim is not to explicitly distinguish demographic features of a subject (e.g., male vs. female). We rather aim at implicitly exploiting such information to improve the accuracy of subject identification. The role demographics plays in authentication has been reported by many recent studies. Exploiting demographic information can entail two possible strategies. Both require pre-determination of relevant demographic classes, that drive the choice of the best suited recognizer in a set of ad-hoc trained ones. In the first strategy, a human operator visually classifies demographic features of the subject to recognize, and runs the appropriate "strong" recognizer. In the second one, the identification of the most appropriate "strong" recognizer follows the results obtained from a set of upstream classifiers for soft biometrics. Both solutions are poorly suited to most real world applications, e.g., video - surveillance. Our architecture mediates recognition across different demographics without any pre-determination of demographic features. We still have different "strong" classifiers, each trained on a demographic class. The probe is submitted to all of them at once. A supervisor module estimates reliability of the single responses, and the most reliable result is returned. In this approach, classifier reliability is not a static feature, but it is estimated for each probe. The proposed multiple-expert system provides similar performance to pre-determination of demographics. Experimental results show higher

✉ Maria De Marsico
demarsico@di.uniroma1.it

[1] Sapienza University of Rome, Rome, Italy

[2] University of Salerno, Salerno, Italy

[3] University of Naples Federico II, Napoli, Italy

[4] George Mason University, Fairfax, Virginia, USA

flexibility, efficacy and interoperability. We also focus on interoperability across face datasets by adopting EGA (Ethnicity, Gender and Age) database as a benchmark, which is obtained by combining images from several publicly available face datasets.

Keywords Biometric systems · Demographics · Face recognition · Interoperability · Soft biometrics · A-posteriori demographics categorization

1 Introduction

Automatic face recognition has been often considered as one of the most successful applications of image analysis and understanding [56]. If one considers the performance claims made over the years, an important fact emerges. Since the beginning, proposed methods seemed to achieve a good accuracy. For instance, regarding the seminal work by Turk and Pentland [53], the authors reported a correct classification rate of up to 96 %, depending on the capture conditions. It is interesting to notice that, in their setting, size was the worst problem, with correct classification of 64 % over size variations. However, as in other fields, as soon as the research finds a satisfying solution to certain problem, attention moves towards new ones to tackle. Good performance achieved in the experiments reported in literature is still hardly reproducible in real world applications, in general, and in uncontrolled settings, in particular. For instance, the degradation of face recognition caused by pose, illumination and expression (PIE) variations continues to spur a huge amount of research. At present, the datasets used as benchmarks contain images that are closer to biometric samples actually captured in everyday applications. Uncontrolled settings are the standard in real life scenarios, e.g., surveillance using smart camera networks, tagging of huge amount of images (e.g., in social networks), and mass security screening. A robust biometric system should show resilience to biometric variability, e.g., incomplete data (e.g., occlusion), corrupt data (low quality or deliberate disguise), as well A-PIE (age-PIE) variability.

Beyond uncontrolled settings, further factors relevant for recognition have recently captured researchers' attention. Among these, the focus of this paper is on the influence of demographics on recognition. Our aim here is not to deal with recognition of such features, but to find a strategy to exploit them implicitly for face recognition, without pre-determining them, neither through human intervention nor through an automatic system. For this reason we omit reviewing the various kinds of approaches addressing demographics recognition; the interested reader can refer to the rich literature about gender recognition (e.g., [34]), age estimation (e.g., [15]), ethnicity recognition (e.g., [18]), and mixed demographics (e.g., [14, 46, 52]). From the point of view of person recognition, demographic features are considered as soft biometrics. In fact, they are not able to identify a subject by themselves. However, they can support different kinds of specific "smart" applications in ambient intelligence, e.g., age-related services. Moreover, they can be exploited in connection with "strong" biometrics as a pre-classification step. Regarding this last point, it is interesting to consider how a preliminary determination of these features can affect recognition. Among the first findings along this line, FRVT 2002 estimated the performance variability for different groups of people, and in particular investigated the effect of demographics on performance [38]. Similarly, gender and aging can hinder recognition, if system training is performed on a dataset where the demographics are not balanced [39, 40]. Therefore, biometric systems trained and tested in a certain country may be less effective on images of a different ethnicity. It is worth noticing that many datasets used for research on face recognition were collected in research centers or universities. As a consequence, they present

not only a prevalence of subjects of the local ethnicity, but also of the age range between 20 and 30. Even the most challenging datasets, e.g., FERET [37], are not always well balanced with respect to different demographic features. When demographic differences are involved, one can further consider the link between biometrics and forensics [20]. The work by Klare et al. [23] is the most recent attempt to systematically investigate the role of cross-demographics in automatic face recognition. The results testify to the fact that a preliminary demographic classification ("binning or stratification") of an image, and its subsequent submission to a specialized classifier (i.e., one trained with face images of that specific demographic group), would improve recognition performance. The implementation of this strategy suggests two possible alternatives. One way to leverage demographics entails the presence of a human operator. The operator visually evaluates the characteristics of the subject to recognize, and submits the probe to the appropriate classifier in a set. In this set, each item is trained on different demographic classes (e.g., a single ethnicity). A possible alternative is to have a set of upstream classifiers, each trained on a specific soft biometrics (in this case demographics, e.g., gender). These would be used in cascade to select a "strong" recognizer, e.g., a face recognizer specialized for young men. In general, the human-in-the-loop approach, i.e, a kind of mixed strategy entailing human feedback during the automatic process, achieves a performance improvement in classification tasks [2]. The continuous presence of a human operator would have additional benefits. These include the possibility to reduce "wild" aspects, by controlling capture conditions such as pose, illumination and expression (PIE). However, this approach is not always feasible or even desirable. On the other hand, the presence of a cascade of biometric systems that gates the image to the corresponding demographic classifier would slow down recognition. Furthermore, it would also be error-prone, therefore introducing an (additional) early error in the chain which would compromise the whole following process. As a consequence, both solutions are poorly suited to unattended settings, and to real world applications requiring a quick response. Our proposal avoids any kind of explicit pre-selection. We still exploit a multi-expert architecture composed of a number of different "strong" classifiers, i.e., face classifiers. Each of them is trained on face images belonging to a certain (combination of) demographic group, and relies on its own gallery of enrolled subjects. Galleries possibly overlap, according to the sets of demographic features, or combinations, that are relevant for the application at hand. The response from each classifier (recognized identity) is augmented with an estimate of its reliability. The latter is not a static feature of each system, but is dynamically computed for each probe. Each probe is submitted to all the classifiers at the same time, and the most reliable response (identity) is returned as the final one. In this way, the proposed approach exploits implicit demographic information coming from a specific training, without determining it a-priori for each single probe. Actually, if also needed, demographic information can be deduced a-posteriori from the source of the selected response: the subject is deemed to belong to the demographic group on which the winning classifier had been trained.

Before going further, it is worth underlining the difference between the problem dealt with by this paper, and the nowadays popular topic of automatic feature selection. In most works along this line, "features" are intended as specific low level ones, e.g., geometrical, texture, or topological characteristics, either local or global, or transformation coefficients, that can improve classification performance. For instance, the work in [24] and [25] investigates the problem of selecting proper Discrete Cosine Transform (DCT) coefficients, to obtain the best discrimination effect in face and palmprint recognition. In many cases the identified features may even lack a precise human "semantic" interpretation. For instance,

this often happens in deep learning [49, 50]. Starting from this low level, our work rather deals with meta-characteristics, high level features that have a precise meaning for human operators too. As such, they can improve the performance of any recognition method, including those based on low feature selection. Of course, it is almost trivial to state that a specialized classifier can do better than a general-purpose one. However, two issues raise regarding automatic face recognition. The first one concerns the automatic choice of the right classifier, that is not always easy or even possible to achieve, unless we know in advance the specialization the sample belongs to. The second and even more interesting one is that, regarding biometrics, it is not completely clear which is the most beneficial information to exploit to improve classification. The research line this paper follows regards confirming the hypothesis that demographics is an effective group of features to increase face recognition performance.

An additional problem that has recently caught research attention is interoperability across different datasets [10]. Many experiments presented in literature use quite similar capture conditions and/or devices for both gallery and probe images. On the other hand, interoperability is the ability of diverse equipment, software applications as well as human stakeholders to work together (inter-operate). It impacts overall system performance and affects the concrete ability to exchange information. Possible factors that affect interoperability include the accuracy of sensors, e.g., image resolution or sensor technology (see for example [45] for sensor interoperability in the case of fingerprints), the effects of lossy image compression, and the acquisition environment [30] (e.g., indoor vs. outdoor). Interoperability requires well defined software/hardware protocol interfaces. These should allow connecting different products and systems (as well as organizations, as it is the case with forensics), without any restrictions about access and implementation. As mentioned above, demographics of exploited benchmarks can also affect interoperability of biometric recognition systems, in the sense that a system trained on some demographics may result less effective with a different population. Interoperability issues are addressed here in terms of database sets, taking into account image quality, and the diversity of the population enrolled. Our contribution with respect to the work by Klare et al. [23] is twofold. First, we confirm the results achieved therein in a different experimental setting, and with interoperable data. Second, we go further by avoiding both the human-in-the-loop limitations of their solution, and the increased computational weight of an automatic explicit soft class pre-selection. Our approach makes the use of demographic information automatic as well as implicit.
The outline for the paper is as follows. Section 2 shortly surveys the role of demographic features on human and automatic face recognition. Regarding interoperability, Section 3 discusses the role of interoperability across different datasets used for training and testing, and Section 4 deals with interoperability across demographics and presents experimental results. Section 5 describes the multi-expert architecture we propose to implicitly exploit demographic features. Section 6 presents the comparative results of experiments using different architectural configurations, that show the feasibility and utility of our novel methodology. Finally, Section 7 draws some conclusions and anticipates future work.

2 The role of demographics in recognition

The face is deemed as the most important biometric trait enabling people to recognize their peers. A huge amount of both past and recent research is devoted to investigate how the human recognition abilities can be reproduced by an automatic system (see the recent [31]).

A person observing a face can also easily make subtle yet correct judgments regarding gender, identity, age, and expression/mood. Nevertheless, a number of studies in literature note that this ability decreases, when such categorization regards individuals from a different demographic group. During life, human recognition skills are nurtured by both cultural and cognitive processes. These include the evolution (phylogeny) within a specific ethnic group, but also the growth of a person (ontogeny) within a mixed population, including different ages and genders. The same combination of such processes may also explain negative human performance in face recognition. For instance, the other-race effect causes a poor ability to recognize a subject from a different race. The notable studies by Goldstein [12] give two possible reasons for this: psychosocial, due either to prejudice, unfamiliarity, or other interpersonal reasons; and psychophysical, due to either different amounts of reflectance from different skin colors, causing different amount of perceived details, or to race-related differences in the variability of facial features. A number of researches aimed at assessing the relative importance of these two groups of factors. Related investigations continue catching experts interest. In particular, the study of the coefficients of variation for different facial features in different ethnic groups, seems to indicate that poor cross-race identification is more likely a psychosocial problem. More recent studies ascribe the other-race effect to the so-called "contact hypothesis" [7], which refers to the single individual experience. Even if it cannot completely explain the phenomenon [33], the concept of "familiarity" is often the best explanation for excellent human recognition performance. Ethnicity is not the only demographic feature that may affect recognition. Further studies by Goldstein quantify the role of gender in face recognition [13]. For instance, it appears that in a Japanese population, most women' facial features are more heterogeneous than the men' ones. A similar but less significant difference seems to hold for white women' faces. Last but not least, age heavily affects recognition across time. The example study in [27] investigates this problem through a real passport photo verification task. Two interesting works deal with a meta-analysis, whose aim is to contrast and combine results from different studies about possible sources of bias in face recognition. One of these works discusses own-gender bias [17]. The authors conclude that, whereas the own-race bias is typical of most ethnic groups, the own-gender bias is asymmetrical, as only females seem to suffer from the effect. The second work regards own-age bias (OAB) [42]. In this case the authors conclude that children, younger adults, and elder adults exhibit superior discrimination ability for same-age compared with other-age age faces.

Returning to automatic face recognition, some recent studies aim at quantitatively comparing the performance of both academic and commercial algorithms against human ability. An example is the research presented by O'Toole et al. [35] regarding the ability to recognize faces under illumination variations. The authors report that, on the less distorted image pairs, all the automated algorithms exhibited better performance than humans; for the "difficult" pairs, this result was not generalized and sometime reversed. However, results of similar comparisons are seldom univocal. In a following work focusing on challenging tasks, O'Toole et al. [36] compare face identification by humans and by machines using images taken under a variety of uncontrolled illumination conditions, in both indoor and outdoor settings. Further variations are represented by hair style, facial expression, hats, glasses, even if pairs of compared face images are all taken in frontal view. Notice that this single condition can make matching less difficult. In fact, even human recognition of unfamiliar faces is affected by changes in viewpoint [16], though it seems that this does not hold for familiar faces [4]. In [36] the authors conclude: "More generally, differences between the performance of humans (at their best) and state-of-the-art face recognition algorithms are analogous to differences between humans recognizing familiar versus unfamiliar people".

The aim of the recent "The Good, the Bad, & the Ugly" (GBU) challenge [41] was just to spur investigation about factors affecting automatic recognition across changes in still frontal faces.

Given the above considerations, it is well reasonable to expect that even automatic face recognition can be affected by specific problems, when cross-demographic features are involved. Indeed this has been reported since the first surveys on the subject (see [6]). However, especially for ethnicity, this phenomenon was seldom noticed or systematically observed. The main reason is that biometric systems implemented in a specific country, e.g., Asian, would usually rely on a dataset of voluntary subjects of the same ethnicity, and often of the same age range (about 20- 30 years old). The present work focuses on how demographic information, when appropriately exploited, can positively affect face recognition performance. On the other hand, lack of demographic-based training can produce negative effects. Among demographics, age of the subject repeatedly shows up as one of the most important personal covariates, in terms of face recognition performance. It mostly, but not only, affects intra-subject recognition across time. Quoting the conclusions in [27]: "In the experiments we observed that the difficulty of face recognition algorithms saturated after the age gap is larger than four years (up to ten years)". In [43] similar results are reported, but, on the other hand, "older" people (but the maximum considered age is only 49) are more easily recognized by PCA (yet when the time elapsed between enrollment and testing is short). Gender and ethnicity can rather significantly influence inter-subject recognition. While training on the same class of ages does not make it easier to recognize a person after twenty years, one can show that training on specific gender or ethnicity can improve human identification [23]. Of course, such factors still mutually influence each other. As an example, age marks may change according to race (white elder people tend to have more wrinkles).

As a last note, it is worth noticing that it is still a matter of debate if other biometric traits suffer from the same sensitivity to demographics. Recent studies confirm a difference in fingerprint image quality across age groups, most pronounced with the over-62 age group [32]. In 1892 Galton published a study about the frequency of arches, loops, and whorls pattern types (which will be the basic elements for the precursors of present fingerprint systems) in fingerprints from various races [11]. One of the aims of the study was to confirm the assumption, which seemed obvious in late nineteenth century, that fingerprint patterns were probably inherited, and should therefore correlate with, say, race, ethnicity, and various behavioral characteristics. However, the author had to admit that he found almost no significant variations, except for slightly fewer arches among Jews. This was what the anthropologist Paul Rabinow has called "Galton's regret" [8].

3 Interoperability across databases

Biometric authentication typically involves four distinct operational stages: feature space learning ("training"), gallery enrollment, testing to evaluate the achieved accuracy, and standard operation ("querying"). In typical laboratory settings, training and testing images are disjoint subsets of a set acquired under relatively well-controlled and quite similar conditions. This often implies the same device and/or backgrounds, and the same type of subjects. In general, no standard operation phase is planned, and used datasets make up a kind of closed world. On the other hand, commercial systems intended for use in real "wild" life

applications, are usually trained (and tested) on sets of images, which may have very little in common with data encountered in future standard operation. Hard real life scenarios include mass screening, surveillance using smart camera networks, tagging fostered by social networks, and biometric management of crowds. In most cases they entail unattended operation. In practice, performance for uncontrolled settings is significantly lower than the performance advertised for large scale, but tightly controlled biometrics evaluations. Moreover, during normal operation, ground truth is usually not available, with false positives not disclosed and false negatives not available. The mismatch between controlled and uncontrolled settings falls under interoperability issues, together with the mismatch between the exploited datasets. Interoperability can be further considered as a thread that links biometrics and forensics. It involves distributed data collections and federated identity management systems, and can provide key performance indexes that ultimately affect large scale deployment. A preliminary study [10] evaluated different operational biometric settings for the purpose of interoperability. The evaluation across face space derivation, enrollment, and testing was pursued by varying both the composition (different datasets exploited for training and testing) and the quality of biometric data. In order to test cross-dataset performance, different datasets were used: 1) FEI dataset [51], including 2,800 images of 200 subjects; 2) PUT dataset [22], including almost 10000 high-resolution images of 100 people; 3) Essex [47], including 7900 images of 395 individuals, divided into the two subsets Essex94 and Essex 95 with 24bit color JPEG image format. Images present slight variations in pose and illumination within and across datasets. Table 1 reports the compositions of the datasets used. Different face spaces were considered according to the (PCA and/or LDA) components used. Performance of PCA is enhanced by dropping the top ranked eigenvectors. They encode mainly illumination changes, and therefore are not relevant to the identification task. The basic findings reported vis-à-vis interoperability in identification mode (Tables 2 and 3) indicate that, even for these basic methods, the quality of images during training/learning the face space is much less important than the quality of images during enrollment and testing. Additional insights suggest that it does not make much difference if the face space is derived from biometric data coming from the same dataset source as that used for enrollment and testing. On the contrary, the size of the gallery and query sets does affect performance. In fact, the relatively poor performance with the FEI dataset comes from having only 4 images per subject, while PUT and Essex have 22 and 20 images per subject, respectively. The results reported seem to validate the use of a separate dataset for training; on the other hand, the training data set must be representative enough of the demographics expected during deployment and use. For more details see [10]. Furthermore, a system to be used in real-world settings should support interoperability across demographics.

Table 1 The composition of the dataset exploited to investigate their interoperability

Dataset	Number of Subjects	Number of images per subject	Total number of images
Essex	223	20	4460
Essex94	152	20	3040
Essex95	71	20	1420
FEI	200	4	800
PUT	100	22	2200

Table 2 PCA Identification. PCA $< x >$ corresponds to the PCA algorithm where the face space is composed of the first 100 eigenvectors excluding the $< x >$ first eigenvectors

Training	Enrollment/Testing	PCA 0	PCA 1	PCA 4	PCA 10	Cross-Validation
FEI	**FEI**	39.58 %	38.79 %	48.71 %	60.88 %	4-fold (Training & Enrollment/Testing)
Essex	**Essex**	99.40 %	99.42 %	99.42 %	99.29 %	5-fold (Training & Enrollment/Testing)
Essex94	**Essex94**	99.87 %	99.87 %	99.84 %	99.84 %	5-fold (Training & Enrollment/Testing)
FEI	**PUT**	92.73 %	93.05 %	94.14 %	95.55 %	5-fold (Enrollment/Testing)
PUT	**FEI**	38.75 %	45.38 %	47.75 %	54.75 %	4-fold (Enrollment/Testing)
Essex	**PUT**	93.00 %	92.50 %	91.36 %	90.41 %	5-fold (Enrollment/Testing)
Essex	**FEI**	38.38 %	38.38 %	38.88 %	38.75 %	4-fold (Enrollment/Testing)
FEI	**Essex**	99.46 %	99.44 %	99.42 %	99.44 %	5-fold (Enrollment/Testing)
PUT	**Essex**	99.39 %	99.37 %	99.30 %	99.28 %	5-fold (Enrollment/Testing)
Essex94	**Essex95**	98.38 %	98.24 %	98.45 %	98.10 %	5-fold (Enrollment/Testing)
Essex95	**Essex94**	99.90 %	99.87 %	99.87 %	99.87 %	5-fold (Enrollment/Testing)
FEI/PUT	**Essex**	99.39 %	99.37 %	99.51 %	99.39 %	5-fold (Enrollment/Testing)

Table 3 LDA Identification. LDA $< x >$ corresponds to the LDA algorithm where the first $< x >$ eigenvectors are excluded and the next 100 eigenvectors are included

Training	Enrollment/ Testing	LDA 0	LDA 1	LDA 4	LDA 10	Cross-Validation
FEI	**FEI**	89.71 %	89.58 %	89.83 %	90.13 %	4-fold (Training & Enrollment/Testing)
Essex	**Essex**	99.51 %	99.50 %	99.52 %	99.51 %	5-fold (Training & Enrollment/Testing)
Essex94	**Essex94**	99.88 %	99.88 %	99.86 %	99.86 %	5-fold (Training & Enrollment/Testing)
FEI	**PUT**	99.27 %	99.36 %	99.32 %	99.45 %	5-fold (Enrollment/Testing)
PUT	**FEI**	69.75 %	68.25 %	68.88 %	69.63 %	4-fold (Enrollment/Testing)
Essex	**PUT**	95.50 %	95.36 %	95.50 %	95.09 %	5-fold (Enrollment/Testing)
Essex	**FEI**	51.38 %	51.63 %	51.63 %	51.50 %	4-fold (Enrollment/Testing)
FEI	**Essex**	99.37 %	99.33 %	99.28 %	99.22 %	5-fold (Enrollment/Testing)
PUT	**Essex**	99.42 %	99.44 %	99.39 %	99.39 %	5-fold (Enrollment/Testing)
Essex94	**Essex95**	98.80 %	98.59 %	98.87 %	98.80 %	5-fold (Enrollment/Testing)
Essex95	**Essex94**	99.84 %	99.87 %	99.87 %	99.87 %	5-fold (Enrollment/Testing)
FEI/PUT	**Essex**	99.28 %	99.35 %	99.26 %	99.28 %	5-fold (Enrollment/Testing)

4 Interoperability across demographics

As already pointed out, the work by Klare et al. [23] is among the most recent and comprehensive attempts to investigate the role of cross-demographics in face recognition. The experiments and results reported show a significant influence of demographics discrepancy between training and testing sets on final recognition performance. The proposed solution is to perform a manual pre-determination of relevant demographic features, which would restrict the search space before engaging the automatic recognition process, and allow to exploit the most suited recognizer. However, in some significant scenarios, such as video-surveillance or access control, it might be not possible to rely on continuous human interaction for the above purpose (human-in-the-loop). To address this interoperability aspect we introduce here a novel multi-expert architectural strategy. It avoids any human intervention in pre-selection of subsets of the search space, related to specific demographics. It is worth underlining that this is not affected by the recognition method used. In other words, the approach is independent from the number and kind of classifiers used, that can be chosen without any restriction. Each classifier is trained on a specific (set of) demographic feature(s). We also avoid using a "soft" classifier upstream, playing the role of first determining demographics of each probe, before submitting it to the right expert. This would slow down the recognition process, and provide more information than really needed. Our final goal does not necessarily imply to determine such demographics. In fact, we will distinguish between explicit and implicit use of demographic information. Furthermore, the preliminary demographics recognition, if wrong, introduces an unrecoverable error in the following steps, which will negatively influence the final result. As an alternative to both the human-in-the-loop strategy and the use of soft biometrics, we want to achieve the best recognition results in the most direct way.

In order to test our approach, we needed an experimental setting quite balanced with respect to the distribution of the different demographics. This allowed avoiding the bias deriving from having more training on some features than on others (or on some classes than on others). No present available dataset offers such characteristic together with an appropriate ground truth. Even MORPH [43], which is one of the largest available ones, is not well-suited to our aims, besides requiring a fee even for its non-commercial release. The most populated MORPH-II subset, includes 55608 color images of 13673 subjects between 16 and 99 years old: 47057 images belong to male persons and 8551 to female ones; 42897 images depict black faces, 10736 white, 1753 Hispanic, 160 Asian, 57 Indian and 5 faces are of other ethnicity. The images have either 200×240 or 400×480 pixels resolution. As it can be deduced, the dataset is highly imbalanced towards black male persons, and lacks images of persons below the age of 16. On the other hand, it was specifically collected to investigate cross-age face recognition. Since each of our classifiers is trained on a different demographic set of features, we would have super-trained and under-trained classifiers. While this is normal in real situations, we wanted to start from a "balanced" setting, and to stress the interoperability aspect. Towards that aim, we used EGA (Ethnicity, Gender and Age) v1.0 database [44], which is composed of images selected from six existing, publicly available datasets in order to guarantee a fair balance in the demographic features (ethnicity, gender, age): 1) CASIA-Face V5 [5], includes 2,500 images of 500 subjects, from a single session captured by an USB camera; image resolution is 640×480, 16bit color; faces are captured at different distances and display illumination and pose variations, and eyeglasses; subjects are mostly young and of Eastern ethnicity; 2) FEI dataset [51], includes 2,800 images of 200 subjects, 100 male and 100 female, each with 14 images; color images

resolution is 640 × 480; faces have been acquired on a white background, and belong to subjects whose age ranges from 19 to 40 years old, mostly of Latin ethnicity; 3) the very popular FERET dataset [37], includes 14,126 images of 1,199 subjects acquired in 15 sessions between August 1993 and July 1996; image resolution is 256 × 384 for 8 bit greyscale; faces are categorized in sets (fa, fb, dup I, dup II) according to pose and acquisition period, and present slight variations in illumination and expression; the dataset is heterogeneous with respect to ethnicity, gender and age; 4) FRGC [39], includes 50,000 images captured in 4,003 subject sessions and divided in training and validation partitions; each subject session includes images captured in controlled (four) and non-controlled (two) conditions, and a 3D model; image resolution is 1704× for 24 bit color; most subjects are of Caucasian ethnicity, and the number of subjects of different ethnicity is quite marginal; subjects are mainly concentrated in a same age range (young/adult), while an adequate number of subjects is present for each gender; 5) JAFFE [29], includes 2130 images of 10 subjects, mainly gathered for facial expression analysis; image resolution 256 × 256 for 8bit greyscale; subjects are all female and Japanese; age too seems to be uniform; 6) Indian Face Database [21], includes 40 distinct subjects in frontal pose with eleven different gaze directions for each individual, plus some additional image when available; the dataset is divided into male and female subjects; image size is 640 × 480 pixels, with 256 grey levels per pixel; images are captured on a uniform background, and present four expression variations; all subjects are of Indian ethnicity, with an adequate distribution with respect to gender, but not with respect to age. The overall composition of EGA with respect to demographic features is summarized in Table 4 (y= young, a=adult, m=middle-aged). In summary EGA dataset contains 469 subjects (about 11 % Afro-American, about 24 % Asian, about 34 % Caucasian, about 16 % Indian, and 14 % Latin). We extracted five images for each subject. The dataset is quite well balanced with respect to gender with 52,4 % male and 47,6 % female, and slightly less balanced with respect to age with 32,6 % young, 48,5 % adult and 18,9 % middle-aged. The latter is due to the fact that most available datasets are acquired in academic settings, and this impacts on the composition of the resulting collection. We note here for completeness that, at present, the most suitable dataset to reproduce real world conditions would be Labeled Faces in the Wild (LWF) [19]. However, it is only annotated with subjects' identity, not with demographics. This makes it hard to exploit for our experiments. It is worth noticing that our multi-expert architecture will not require pre-determination of probe demographics, but in the experimental phase this information is used to measure performance and compare different approaches.

Before presenting our proposed method, it is worth establishing a common ground with the work by Klare et al. In a similar way, we also rely on multiple classifiers, each trained on a specific demographic feature. Our preliminary experiments focused on the influence of different training strategies on the final recognition results. In particular, we chose ethnicity as the first test-bed. Compared with gender and age, this feature usually presents the highest number of different significant classes, and therefore is prone to the highest degree of confusion. In fact, gender only entails a binary classification. Regarding age, it is not usually processed neither at single year level, nor in narrow bands, that are very difficult to partition even for a human operator, but rather in quite wide intervals (typically, child, adult, and elder). Being possibly subject to higher error rates, yet allowing a reliable verification, ethnicity is a better testbed to assess the advantages of the proposed approach. We decided to use quite popular and widely available recognition procedures. The aim was to obtain baseline and repeatable results, and to better assess the influence of demographic information. LBP [1] was used to extract a feature vector from an image. In practice this kind of

Table 4 Overall demographic composition of EGA (y=young, a=adult, m=middle-aged).

Afro-American						**Asian**						**Caucasian**						**Indian**						**Latin**					
53						111						162						75						68					
Male			**Female**			**Male**			**Female**			**Male**			**Female**			**Male**			**Female**			**Male**			**Female**		
20			33			54			57			89			73			49			26			34			34		
y	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**	**y**	**a**	**m**
3	13	4	16	11	6	34	14	6	33	19	5	25	50	14	20	33	20	3	37	9	4	15	7	7	19	8	8	16	10

vector is a chain of histograms of LBP codes, computed for each cell of a grid defined over the image. The finer the grid, the larger the vector size. Afterward, OLPP dimensionality reduction method [54] was used to obtain a more compact vector. Of course, other dimensionality reduction methods might have been used, e.g. the approach in [26]. By our experiments, it comes out that the results relevant for this work are not dramatically affected by this choice. In the following we will refer to the obtained classifier as LBP-OLPP. We remind that in EGA each subject has five images. Three images per each subject were used for training, to model the transformation matrix to map the original vectors onto the new space. The remaining two images per subject were used one for the gallery and the other as probe. The first part of the experiments aimed at ensuring that the behavior of the classifier was qualitatively consistent with the outcomes reported in [23], in order to be able later to compare the contribution of our proposal.

The first experiment assessed how and to what extent the ethnic composition of the image set used to train a given classifier could affect its performance, in case the system operated in a different context. For each experimental session, the classifier was trained on one out of five different ethnicities: Afro-Americans (AA), Asians (AS), Caucasians (CC), Indians (IN) and Latinos (LT). Each time, the probe and gallery images included only members of the ethnic group under consideration, while the classifier was trained either on the same or a different ethnic group. The experiment was repeated separately for each of the five ethnic groups. Experiments were carried out according to the following. Recognition was carried out in identification mode, i.e., the identity was searched through the whole gallery, and therefore the probe was matched against each stored template. During testing, the classifier ordered the gallery templates according to the similarity score (Euclidean distance between the feature vectors) obtained by matching each of them with the probe. The first template in this ordered list was assumed to provide the probe identity. We adopted a mixed identification strategy. As for the composition of the probe set, we exploited a closed-set modality (each probe belongs to a subject which is present in the gallery). However, in order to accept the result, the corresponding score was compared with an acceptance threshold *th*, i.e., like in open-set modality, the first retrieved identity was returned as accepted only if score was higher than *th*. This seemed more realistic than classical closed-set modality, when the first identity is accepted in any case. Denoting the probe identity as *id(p)*, and the identity of the first returned subject as *id(g)*, (*id(p)* =*id(g)*) holds when the identity of the first returned template is the same as the identity of the probe. We have the following four possible cases:

1. Correct Identification : Correct User ($id(p) = id(g)$) and Accepted ($score(g) > th$) (CI);
2. False Reject: Correct User ($id(p) = id(g)$) and Not Accepted ($score < th$) (FR);
3. Correct Reject: Incorrect User ($id(p) \neq id(g)$) and Not Accepted ($score < th$) (CR);
4. False Identification: Incorrect User ($id(p) \neq id(g)$) and Accepted ($score > th$) (FI)

Cases CI e CR represent safe system behaviors, FR can be disturbing for the user, while the case to avoid is FI. System performance can be measured according to the corresponding rates, i.e. Correct Identification Rate (CIR - correct identifications vs. the cardinality of the set of probes), False Reject Rate (FRR - false rejects vs. the cardinality of the set of probes), Correct Reject Rate (CRR - correct rejects vs. the cardinality of the set of probes), and False Identification Rate (FIR - false identifications vs. the cardinality of the set of probes). Notice that CIR+FRR+CRR+FIR=1. It is possible to study the system performance at different operation thresholds. Since it is the most critical value, we fixed FIR and determined *th* accordingly. In our experiments, we set *th* such that FIR=$10^{-1} = 0.1$, and computed the other values using this threshold.

Table 5 shows performance of the LBP- OLPP classifier with different, single-ethnicity training sets, when applied to probe / gallery sets belonging to a single ethnicity as well. We remind that AA stands for Afro-Americans, AS for Asians, CC for Caucasians, IN for Indians, and LT for Latinos.

The results in Table 5 confirm the intuitive hypothesis that a classifier trained on images of a specific ethnic group gives the best results when the probe / gallery images belong to the same ethnic group. The only exception is observed when the classifier is trained with Asian ethnicity. In that case the results are comparably good. Additional tests reveal that some differences in performance can be attributed to the different number of subjects available for each ethnic group. It is worth reminding that the images in the EGA database come from different datasets, therefore they are captured under different illuminations and different resolutions. LBP-OLPP is quite robust to these factors. The behavior emerging from Table 5 is consistent with results reported by Klare et al.

We also tested the classifier when it was trained on a single ethnicity but tested on probe / gallery sets with images from all ethnic groups (all five in this case). Table 6 shows the results.

We can observe that CIR is always lower than that achieved when the classifier is trained and tested on the same ethnicity. It is worth noticing that this is the case when a system is trained within a specific ethnic setting (e.g., university laboratory, or software company) but then used in a more general context. Again, these results are consistent with those by Klare et al. The opposite situation is less interesting in practice, since in real world applications it is more likely that the classifier is trained on a set of images that is less representative and varied than that processed at operating time. However, for sake of completeness, we also measured performance when the classifier was trained on images from all ethnic groups, and then tested on a single one. Table 6 shows this result too.

Note that most existing biometric systems do not make a preliminary judgment about the ethnicity of a probe image. For this reason, the most realistic situation is one in which the classifier is trained on all ethnic groups and is then tested on all ethnicities. LBP-OLPP achieved a CIR of 0.68 in this case, which is well below the performance achieved by the same classifier when applied to single ethnic groups. The advice than, consistently with [23], is to exploit information from single demographics to improve performance. Our proposal develops along this line and is detailed next.

5 System architecture

5.1 Framework overview

The results discussed in the above section confirm that a classifier trained on a particular value for a demographic feature, can provide optimal performance when tested on samples with the same value for that feature. Starting from this, the first step to carry out is to divide the entire population according to ethnicity, gender and age. The different categories are clearly not disjoint. For instance, a subject can be both young and Afro-American, and thus belong to both categories. Five groups are considered for ethnicity (Afro-Americans, Asians, Caucasians, Indians, Latins), two groups for the gender (males, females) and three groups for age (young, middle-aged, adults). After this, it is necessary to train a number of different classifiers (experts) on different demographic super-classes. These super-classes may represent mixtures of different features, e.g. Asian males. The last step is to devise how to use such information during testing, i.e. during recognition of probes. As already

Table 5 GAR of LBP-OLPP trained on single ethnicities (rows) and tested on single ethnicities with FIR=0.1

	AA			AS			CC			IN			LT		
	CIR	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**
AA	0.717	0.126	0.056	0.698	0.126	0.075	0.698	0.126	0.0754	0.641	0.201	0.566	0.660	0.164	0.0740
AS	0.813	0.068	0.0165	0.831	0.058	0.009	0.822	0.068	0.009	0.803	0.078	0.028	0.831	0.068	0.012
CC	0.722	0.091	0.082	0.765	0.085	0.0457	0.802	0.060	0.035	0.765	0.091	0.042	0.778	0.072	0.048
IN	0.743	0.102	0.054	0.797	0.089	0.013	0.784	0.089	0.027	0.837	0.063	0.000	0.770	0.102	0.021
LT	0.765	0.912	0.032	0.794	0.076	0.294	0.735	0.105	0.047	0.720	0.0911	0.076	0.808	0.076	0.014

Table 6 CIR/CRR/FRR at FIR 10^{-1} of LBP-OLPP classifier when trained on single ethnicities and tested on probe/galleries with five ethnicities (first row) or when trained on five ethnicities and tested on probe/galleries with single ethnicities

Training / Testing	**AA**			**AS**			**CC**			**IN**			**LT**		
	CIR	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**
Single/All	0.641	0.183	0.064	0.822	0.068	0.009	0.790	0.067	0.043	0.824	0.075	0.001	0.720	0.120	0.035
All/Single	0.698	0.126	0.061	0.822	0.068	0.09	0.802	0.060	0.033	0.824	0.076	0.001	0.764	0.091	0.044

discussed above, it is possible to use it explicitly or implicitly. In the former case, it is required to set up a workflow wherein demographic features are determined in advance, in order to select the most appropriate classifier (expert) accordingly. Only this classifier will carry out recognition. In turn, the preliminary identification of the demographic class can follow two modalities. In the first one, it is carried out manually by a human operator: we will refer to this modality as Manual Explicit Managing of Demographic (MEMoD). A second alternative for explicit use of demographics is the automatic selection of the classifier by one or more upstream Demographics Modules (DMs). The probe is first submitted to these modules, which determine its demographic class. In this way the corresponding classifier can be called for recognition. We will refer to this modality as Automatic Explicit Managing of Demographic (AEMoD). Figure 1 shows a schema of MEMoD and AEMoD.

The third possibility, and the one whose implementation is proposed here, is to use demographic information implicitly, i.e., without having to determine it before recognition. This approach exploits a multi-expert architecture ("gated network"): the probe is submitted to all the pre-trained classifiers at the same time, and each of them returns a possible result. A Supervisor Module (SM) implements an appropriate policy to determine the final response. This modality will be referred as Implicit Managing of Demographics (IMoD). IMoD can be configured either with overlapping, or with non/overlapping galleries (discussed in further detail later on) (see Fig. 2). It is to point out that SM is a light module, which requires a computational time not comparable with that required by a DM. The former has only to analyze and combine the responses produced by the classifiers, while the latter has to perform a possibly complex processing to determine the demographic class of the probe. In order to feed the SM, expert modules involved in IMoD return both a list of ordered similarity scores, the first of which will correspond as usual to the identified identity, and also a reliability measure for each response. It is worth underlining that we are not interested in determining the value of each demographic feature for each probe, but rather in avoiding

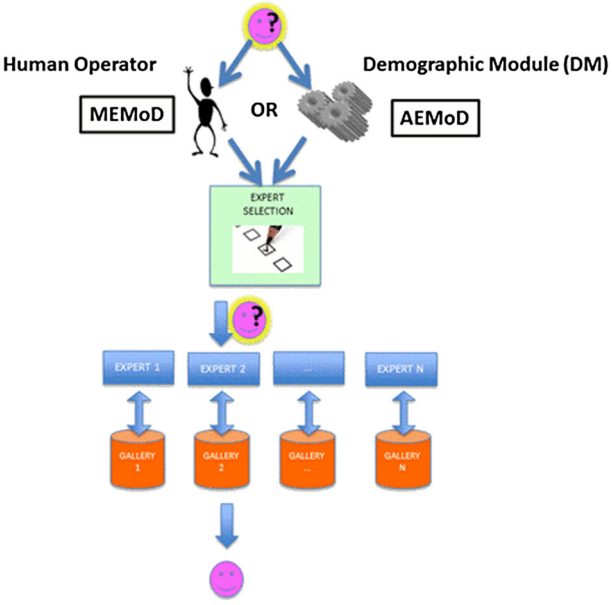


Fig. 1 Explicit use of demographics: MEMoD and AEMoD modalities

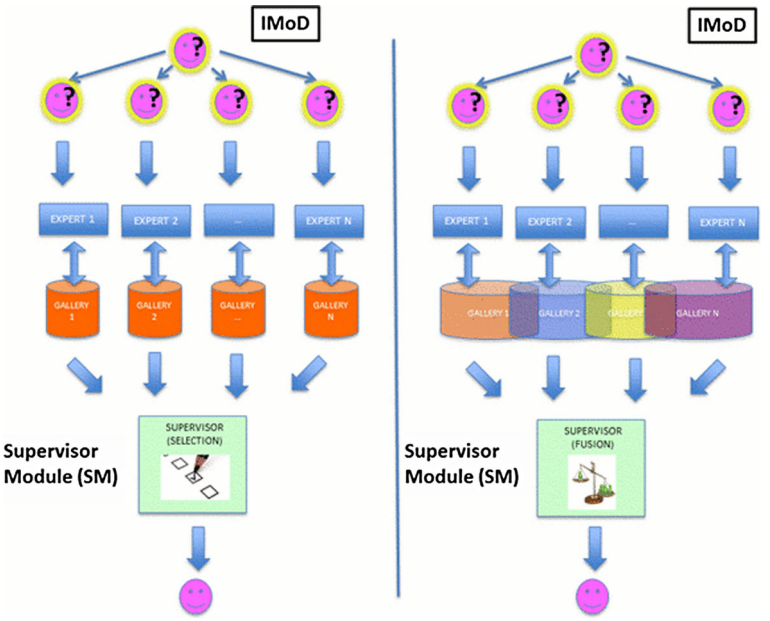


Fig. 2 Implicit use of demographics: IMoD modality (with partition or division strategy for gallery templates)

that the lack of demographic information can possibly negatively affect the final recognition result. In a different perspective, we aim at exploiting such information by adopting IMoD without pre-determining it explicitly for each probe, and obtain the same results as it was explicitly known before starting the recognition process (MEMoD or AMoD). At the same time, our complementary goal is that, during normal system operations, an operator's check may be only required in particularly hard cases, appropriately signaled by an alert system.

5.2 Possible implementations

This section will briefly sketch the lines for the implementation of the different modalities mentioned above, and then the following ones will describe the kinds of modules involved. a) MEMoD (manual selection of the expert): the selection by a human operator is simply simulated by parsing the file names in EGA. As a matter of fact such names are composed by the ordered sequence of the values of each demographic feature for that image (see [44]). b) AEDoG (automatic selection of the expert): a different AdaBoost classifier was trained for each demographic group (see section below for details). Its response is a value w which indicates the confidence of probe belonging to a specific group. Were we purely interested in the demographic information, this value might be compared against a classifier-specific threshold defined during training, so that a 0/1 response is returned. For our tests, we do not use the threshold but rather maintain the returned w value. When a probe is submitted for recognition, we pass it to the Demographic Module (DM), which runs in parallel N demographic classifiers, one for each demographic class, each corresponding to one of the trained face recognizers running downstream. The DM returns the vector of w values produced. The probe is assigned to the class for which the DM has returned the highest w, and therefore submitted to the corresponding expert for recognition. c) IDoG (implicit use of

demographic information): the probe is submitted to all experts at the same time, without any starting hypothesis. Each classifier computes an ordered list of scores for the templates of its own gallery, and passes such list to the Supervisor Module (SM). The critical aspect in this modality is the fusion strategy adopted by the SM to integrate the responses from the experts. It is worth underlining that, though the presented experiments entail the same classifier/recognition algorithm for each class of demographics, this is not a constraint for the system. As a matter of fact, since the experts are loosely coupled through the supervisor, they can implement different classifiers/recognizers (possibly off-the-shelf), especially chosen according to their possible higher effectiveness on a specific demographic feature. We chose to use the same basic feature extraction, both for demographic feature classification and face recognition, because our goal was to demonstrate that a system with the proposed architecture can provide good performance even without demographics pre-detection. For the same reason, the chosen algorithm run from the single classifiers/recognizers is not the most advanced in literature.

5.3 The demographics module

Face demographic categorization can be considered as a preliminary step for recognition, as it allows a reduction of the search space. For instance, it allows to compare the facial template of a middle-aged Afro-American male only with those templates in the gallery that belong to the same category. Humans are able to exceed the performance offered by automated systems, just because they naturally rely on such auxiliary information. However, making this process automatic is in itself an active field of research [55]. Some critical aspects that must be taken into account are, for example, the type of categorization (*hard*, if the subject is associated with a single category, or *soft*, if the subject may belong to more than one category with different confidence), the error introduced by the categorization system (a template can be assigned to the wrong category, thus affecting the performance of the recognition system downstream), the specialization of the classifiers with respect to the characteristics of each category. Sun et al. [48] use a genetic algorithm to select the features that were specific for the discrimination of the gender from the face. The features on which the genetic algorithm works are those provided by the application of eigenfaces to face images. The tested classifiers are a Bayesian one, a neural network, a Support Vector Machine, and one based on LDA. The authors of [28] rather focus their attention on ethnicity. They divide the population into two groups, Asians and non-Asians. Through LDA, the faces are mapped onto a new feature space, and the distribution of each of the two groups is approximated with a single Gaussian model (a Gaussian function, for which one estimates the parameters). The category to which a new sample must be associated can be determined by Maximum a Posteriori Probability (MAP). The work in [18] makes a finer partitioning of the population into three categories: Asian, European and African. For each category, the features are extracted by applying the Gabor Wavelet Transform with retina-sampling. Classification is based on a Support Vector Machine. The determination of ethnicity in this case is in cascade: given an input sample, it is decided whether it is African or not, and in the negative case if it is Asian or not (and therefore European).

In our work, demographic classifiers were trained according to the demographic subsets used for the experiments. They made up the Demographic Module (DM). Section 6 reports details on the different subdivision strategies adopted. As mentioned above, LBP [1] was used to first extract a feature vector from an image, independently from the demographics to be coded. The space of the characteristics of each demographic category was then modeled applying the OLPP dimensionality reduction method [54]. Training aimed at modeling the

transformation matrix towards the lower dimensionality space. This correspond to what we mention as LBP-OLPP classifier. Of course, training is carried out with different training set according to the demographic class. An alternative choice would have been to select a different appropriate feature set for each demographics, since a generic one might be effective only for some of them. The choice of always using the same aimed at maintaining the most basic conditions. For each category c_i in its own new feature space, an ensemble T_i of classifiers was trained, according to the AdaBoost scheme. For the purpose of training, a single image for each identity was considered, in agreement with the statement made by Bruyer et al. in [3]. The authors observe how it is difficult, if not impossible, to separate the information related to the identity of a subject from the peculiarities by which she/he shows the demographic traits of her/his category. However, a classifier T_i in our implementation does not return a 0/1 answer, as it is commonly the case, but it is modified in such a way that it will respond with a weight w_i, which indicates the degree of membership of a face to the category corresponding to T_i. It is therefore not necessary to organize the operation of demographic detection in cascade, as was done in [18], and the face is submitted in parallel to all the classifiers T_i: the category assigned to the face corresponds to that of the T_i which yields the highest w_i. This is also the category of the face recognizer launched downstream.

5.4 The expert modules

Each expert module implements a classifier (face recognizer) that is independently trained on samples that share a common value for one or more demographic features. Each expert is of course trained on a data set with no overlap with data sets used later for enrollment or testing. This avoids the bias related to a lack of generalization ability. The number and type of experts strongly depends on the identified relevant demographic features and on how they are used to partition the whole dataset. However, thanks to mutual complete independence, the architecture is flexible enough to allow the dynamic introduction and/or deletion of experts. Here we consider three demographic features: ethnicity, gender and age. A number of different strategies exist, according to which a population can be divided in subgroups and then assigned to a specific expert. In particular, we examined two possibilities, which seemed the most intuitive ones. In the first case, a subgroup is all-inclusive and membership is specified by a value for each relevant feature. For instance, “Afro-American Male Young” would include all the samples of young Afro-American boys. In this case a partitioning would be produced, since this subdivision entails non-overlapping groups, whose union makes up the whole population. In the second case, the subdivision is based on overlapping (in terms of subjects) subgroups. In practice, a subgroup is specified here by the value for a specific feature, and a subject membership holds if the subject presents that value for that feature. For instance, a subject could be both Afro-American and Young, and therefore would belong to both the Afro-American and Young overlapping subgroups. More precisely, a subject would belong to one and only one subgroup for each demographic feature. In the present case, this entails belonging to three subgroups (one out of five for ethnicity, one out of two for gender, and one out of three for age). The way the whole population is subdivided is important, since it determines different possible strategies for the Supervisor Module. From an implementation point of view, each expert has its own gallery (enrolled subjects). These are drawn from the class of faces it is trained to recognize, according to the chosen demographic subdivision of the population. When a face (probe or test sample) is submitted to the expert, the expert compares it with those in the gallery, and returns a list of gallery subjects ordered by similarity measure. Each similarity measure indicates how much the probe is similar to the corresponding face in the gallery. It is worth noticing

that it is possible to use any off-the-shelf face recognizer, given that it returns a list of identities ordered by similarity with the probe. As already mentioned above, the experiments presented here exploited quite basic techniques, to better stress the influence of demographic information. Even for the expert module, we used LBP to extract a feature vector from an image, and OLPP dimensionality reduction to obtain a more compact vector; gallery images were divided as in Section 4 (three for training, one for gallery and one as probe).

5.5 The supervisor module

For sake of readers, it is worth reminding that a Supervisor Module (SM) is involved only in IDoG (implicit use of demographic information), where no pre-determination of demographic information is carried out (no DM is involved). SM receives a list of candidate subjects as response from each face recognizer, ordered by similarity with the probe p. For each such list, SM first computes a reliability index, which represents how much the potential response provided by the classifier (the first element in the list) can be trusted. A high (low) similarity for the first subject, joined with a high reliability, indicates that the expert is quite sure that the subject is (not) in its gallery. A high (low) similarity, yet with a low reliability value, indicates that the expert believes to have (not) found the identity for the probe within its gallery, but also alerts the system of a high probability of error. The similarity index is a numeric value in the interval [0, 1], while the reliability index ϕ is defined as in [9]:

$$\phi(p) = 1 - |N_b|/|G| \tag{1}$$

with

$$N_b = g_{i_k} \in G | F(d(p, g_{i_k})) < 2 \cdot F(d(p, g_{i_1})) \tag{2}$$

where $|G|$ is the gallery size, p is a probe template, d is any distance measure used in the system to evaluate template similarity, F is a normalization function (we used the Quasi-Linear Sigmoidal (QLS) in [9]), and g_{i_k} is a gallery element (template), with i_k indicating its position in the similarity ordered list, so that g_{i_1} is the first element in the list corresponding to identification matches for p.

The supervisor module collects the responses from the single experts and integrates them according to the data fusion strategy adopted. The fusion policy implemented is strongly influenced by two factors: a) the way the whole population is subdivided into subgroups; and b) the fusion rule. When subgroups are a partition of the whole population, the galleries of the single experts are disjoint. Therefore, the candidate identities returned by each expert are all different from each other. In this scenario, the supervisor module does not (and could not) integrate the responses, but only acts as a selector. In other terms, it must simply choose the correct response from the set of responses by the different experts. Different rules are possible. In the present experiments, SM uses a combination of similarity and reliability indices, which accompany the response of each expert. For each response, the SM computes their product cl, and selects the identity returned by the expert that yielded the maximum for cl. An alternative strategy would set a reliability threshold th_j for each expert, and would select the response with higher similarity score, only among those whose reliability index is above the corresponding threshold. It is interesting to note that, in this case, with respect to the single (disjoint) demographic groups, the performance of the global system cannot be more accurate than the performance of the relevant expert, since this is the only one to store the correct identity in its gallery.

If the population is divided into non-disjoint groups, the supervisor can adopt a policy to integrate responses based, say, on a majority rule. In fact, the same subject belongs to

more groups (e.g., a young Asian male belongs to three groups: Asians, Young and Male groups). In this scenario, the supervisor can select the identity that received more non-contradicting votes (was returned by more non-contradicting experts), if one exists. Only if this does not happen, the supervisor uses a fusion rule for selection, and returns the identity with higher potential in terms of similarity with respect to the reference gallery (similarity index, reliability index, or a combination). Figure 2 shows the two gallery settings. The experiments presented here exploit the same choice rule described above.

6 Experimental results

Performance of the proposed IMoD strategy was evaluated within the same application scenario described in Sections 3 and 4, and compared with systems exploiting different approaches. The population examined is composed by the subjects included in EGA. The demographic features relate to ethnicity (5 categories), gender (2 categories), and age (3 categories). We report performance measured in terms of CIR, CRR, FRR at an operational threshold determined by setting again FIR=0.1. In particular, for each experiment, the performance of three different systems was compared. The first one was MEMoD (see Fig. 1), and exploited the naming convention used by EGA files in order to automatically select the class of the probe and submit it to the right classifier(s). This pre-selection is similar to Klare et al. approach, where this task is assigned to a human operator, and is potentially the one achieving the best performance. The second system implemented AEMoD (see Fig. 1): a Demographic Module performs pre-selection of the appropriate classifier, with more computational demand and possible errors. Finally, the performance of the proposed IMoD (see Fig. 2) was evaluated, based on the multi-expert strategy. Experts were trained on a set of images consistent with the composition of their gallery (see below). All the experiments used LBP-OLPP, for both demographic classification and face recognition.

To better evaluate the results, the performance of each single expert is presented first, referring to the subset of the probes that actually correspond to identities in its reference group. This aims at assessing the performance when recognizing samples that the classifier is trained for. In particular, we consider separate experts for the five ethnicities, for the two genders and for the three ranges of age. Results are in Table 7. As for ethnicity, it is possible to observe that the Afro-American ethnicity gives lower results than the others. The experiment is interesting. Ascribing this lower value to the quality of images (illumination or resolution) would contrast with the fact that the images of other ethnicities, such as Asians, present even worse variations, without the same decrease in performance. Since the classifier was always the same, it is interesting asking if it is reasonable to conjecture a higher recognition difficulty for this ethnicity. Performance for gender was well balanced, and this may mean that no gender category is more difficult than the other. As for age, the LBP-OLPP achieved significantly lower performance on category “Adult” that on the others. However, this category is also the richer in samples (227), compared to 153 Middle-Aged and 89 Young subjects.

In the first comparative experiment, the whole population was partitioned into disjoint groups according to a single demographic feature. The first partition considered entails five groups according to different ethnicities. Therefore, five face recognizers were trained, one for each ethnicity. The Demographic Module involved in AEMoD included five demographic classifiers, and IMoD included its Supervisor Module. The same experiment was repeated by partitioning according to the gender, and to the age , with suitable consistent variations in the combinations of modules. Table 8 reports the compound results of the three

Table 7 CIR/CRR/FRR at FIR 10^{-1} from the five experts trained on ethnicity, from two experts trained on gender, and from three experts trained on age

		CIR	CRR	FRR			CIR	CRR	FRR
ETHNICITY	Afro-American	0.717	0.126	0.056	**GENDER**	Male	0.632	0.080	0.185
	Asian	0.831	0.058	0.009		Female	0.744	0.046	0.097
	Caucasian	0.802	0.060	0.035	**AGE**	Young	0.799	0.061	0.027
	Indian	0.837	0.063	0.000		Medium Aged	0.774	0.037	0.084
	Latin	0.808	0.076	0.014		Adult	0.584	0.102	0.181

partitions (over either five ethnicities, two genders, or three age ranges) with the different strategies. These results suggest, as in [23], that if it was possible to identify a-priori the ethnicity of the subject and to submit the image to the classifier trained for that specific ethnicity, the accuracy would be greater. Some further observations are worth about the results in Table 8. In all three cases AEMoD is closer to MEMoD than IMoD, which however is often the one with best performance. This can be explained by considering that the reliability index used by the SM is an additional information, and concretely influences the choice of the returned identity. Moreover, all three systems behave in a consistent and comparable way with respect to the three features, e.g., they all provide the best performance on ethnicity.

In the second experiment the population was divided in 15 disjoint groups corresponding to 5 possible ethnicities by three possible ages, so that the Supervisor in IMoD acted as a selector. In particular, it implemented the first fusion strategy described in Section 5.4. Table 9 compares performance. Left part of Table 9 shows that both automatic systems decrease their performance with respect to Table 8, due to the high number of subgroups resulting from the combination of different demographics. However, IMoD is more seriously affected. This has a rationale. If we consider a higher number of subgroups, we have a lower number of samples in each subgroup, which in a manual pre-selection would always be assigned to the right expert. In IMoD, high fragmentation is detrimental, since the selection by SM must rely on a single choice in a high number of responses, with the consequence of a higher probability of an error.

In the third experiment we had overlapping non-disjoint groups, and the Supervisor Module implemented the second fusion strategy described in Section 5.4. The whole population was divided into 10 groups corresponding to 5 ethnicities, 2 genders, and 3 ages. Notice that the number of groups for each demographics is just summed up, since they were not a partition. In other words, the groups overlapped with each other, since we had a full partition for each single feature, but each element of such partition included subjects with different

Table 8 CIR/CRR/FRR at FIR 10^{-1} from different systems when the population is partitioned according to a single demographic feature

	Ethnicity			Gender			Age		
	CIR	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**
MEMoD	0.81	0.07	0.02	0.67	0.08	0.15	0.67	0.07	0.16
AEMoD	0.81	0.07	0.02	0.67	0.08	0.15	0.62	0.10	0.18
IMoD	0.79	0.09	0.02	0.78	0.10	0.02	0.79	0.09	0.02

Table 9 CIR/CRR/FRR at FIR 10^{-1} from different systems when the population is either partitioned into 15 disjoint subgroups or divided into 10 overlapping subgroups

	15 disjoint subgroups			**10 overlapping subgroups**		
	CIR	**CRR**	**FRR**	**CIR**	**CRR**	**FRR**
MEMoD	0.814	0.036	0.045	0.741	0.070	0.113
AEMoD	0.710	0.076	0.113	0.702	0.068	0.122
IMoD	0.601	0.163	0.126	0.782	0.085	0.027

values for the other features. The SM returned the most voted identity or, when all received only one vote, returned the identity with the maximum product between similarity and reliability. Results are in the right part of Table 9. Performance of both MEMoD and AEMoD decrease with respect to left part of Table 9 (partitioning). This happens because the category assignment is performed a-priori, therefore the system does not exploit the redundancy (overlap) among the groups. On the contrary, in the IMoD system, the SM can combine, and not only select, the different responses, and also exploit the majority vote as a further confirmation of the correctness of the selected identity to return. Results (see Tables 8 and 9) seem to indicate that overall performance is lower than with demographically homogeneous datasets, therefore confirming the results by Klare et al. We can further observe that IMoD performance, tough slightly lower than in Tables 8, does not degrade by much, while it is better in terms of feasibility / deployment and computation.

7 Conclusions

Current literature suggests that, when demographics are relevant discriminating features in a population, the pre-selection of an appropriately trained recognition system can significantly improve performance. Pre-selection entails either human intervention, or that a demographic module (soft biometrics) runs before the recognition process, in order to trigger the appropriate classifier. The first solution is not always applicable, e.g., in unattended mass screening biometric recognition applications, like tagging of huge amounts of images. The use of soft biometrics may slow down the overall process, since two different classifiers must be run in sequence, and also introduce an early error affecting the overall recognition procedure. This paper describes a novel system architecture to handle recognition across different demographic groups. It uses demographic information, yet implicitly. We avoid determining the value for demographic features for each single probe, thus avoiding both human intervention and the increase of computational time. We use a multiple-expert approach, with different classifiers trained on different demographic classes and run in parallel on each probe. Thanks to this, the only overhead is caused by a supervisor module, which only executes a couple of simple mathematical operations to fuse the results returned by the single experts. As expected, this strategy may achieve slightly lower performance, since the choice among more responses from different experts can introduce further error in the final result. To confirm this influence, we can observe that the smaller the subgroups (the higher the fragmentation, and therefore the number of experts, but also the higher the possible confusion), the worse the results. However, the positive counterpart is the absence

of both manual and automatic pre-processing step in the recognition workflow. Future work will investigate more fusion strategies, to better exploit the information overlap provided by non-disjoint galleries. It is also planned to carry out experiments using the LWF dataset. To this aim, and to avoid an often unfeasible preliminary classification of enrolled subjects, we will also test our architecture with experts trained on different classes, but working with the same overall gallery.

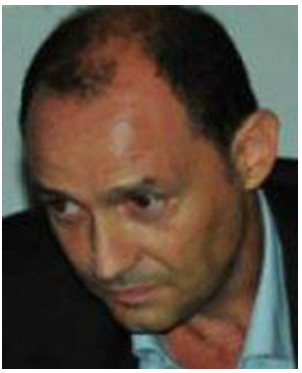
References

1. Ahonen T, Hadid A, Pietikinen M (2004) Face recognition with local binary patterns. In: Proceedings of European Conference on Computer Vision - ECCV 2004, pp 469–481
2. Branson S, Van Horn G, Wah C, Perona P, Belongie S (2014) The ignorant led by the blind: a hybrid Human-Machine vision system for Fine-Grained categorization. Int J Comput Vis 108(1-2):3–29
3. Bruyer R, Leclere S, Quinet P (2004) Ethnic categorisation of faces is not independent of face identity. Perception 33(2):169–180
4. Burton AM, Jenkins R, Schweinberger SR (2011) Mental representations of familiar faces. Br J Psychol 102:943–958
5. CASIA-FaceV5, Available at http://biometrics.idealtest.org/. Accessed January 2015
6. Chellappa R, Wilson CL, Sirohey S (1995) Human and machine recognition of faces: a survey. Proc IEEE 83(5):705–740
7. Chiroro P, Valentine T (1995) An investigation of the contact hypothesis of the own-race bias in face recognition. Q J Exp Psychol Hum Exp Psychol 48A:879–894
8. Cole SA (2004) Fingerprint identification and the criminal justice system: historical lessons for the DNA debate, in David Lazer (Ed.), DNA And The Criminal Justice System - The Technology of Justice (pp. 63–89) MIT Press
9. De Marsico M, Nappi M, Riccio D, Tortora G (2011) NABS: Novel Approaches for biometric systems. IEEE Trans Syst Man Cybern Part C Appl Rev 41(4):481–493
10. El Khiyari H, De Marsico M, Abate AF, Wechsler H (2012) Biometric Interoperability Across Training, Enrollment, and Testing for Face Authentication. In: Proceedings of 2012 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (bioMS 2012), Salerno (Italy),September 14 2012, pp 1–8
11. Galton F (1892) Finger prints. MacMillan and Co., New York
12. Goldstein AG (1979) Race-related variation of facial features: Anthropometric Data I. Bull Psychonomic Soc 13:187–190
13. Goldstein AG (1979) Facial feature variation: Anthropometric data II. Bull Psychonomic Soc 13:191–193
14. Gutta S, Huang JRJ, Jonathon P, Wechsler H (2000) Mixture of experts for classification of gender, ethnic origin, and pose of human faces. IEEE Trans Neural Netw 11(4):948–960
15. Han H, Otto C, Jain AK (2013) Age estimation from face images: Human vs. machine performance. In: Proceedings of 2013 International Conference on Biometrics (ICB), vol 8, p 1
16. Hancock PJB, Bruce V, Burton AM (2000) Recognition of unfamiliar faces. Trends Cognitive Sci 4:330–337
17. Herlitz A, Lovén J (2013) Sex differences and the own-gender bias in face recognition A meta-analytic review. Vis Cogn 21(9-10):1306–1336
18. Hosoi S, Takikawa E, Kawade M (2004) Ethnicity estimation with facial images. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition 2004. IEEE, pp 195–200
19. Huang GB, Ramesh M, Berg T, Learned-Miller E Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, University of Massachusetts, Amherst, Technical Report 07-49, October (2007). Available online at http://vis-www.cs.umass.edu/lfw/lfw.pdf. Accessed January 2015
20. Jain AK, Klare B, Park U (2011) Face recognition: Some challenges in forensics
21. Jain V, Mukherjee A The Indian Face Database (2002). Available online at http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/. Accessed January 2015
22. Kasiński A., Florek A, Schmidt A (2008) The PUT database. Image Processing & Communications 13(3-4):59–64

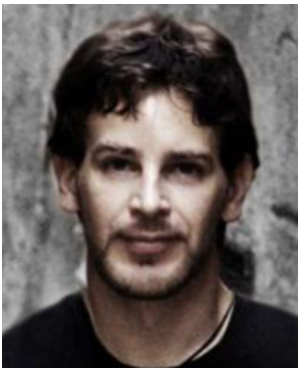
23. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK (2012) Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security 7(6):1789–1801
24. Leng L, Zhang J, Khan MK, Chen X, Alghathbar K (2010) Dynamic weighted discrimination power analysis: a novel approach for face and palmprint recognition in DCT domain. International Journal of Physical Sciences 5(17):2543–2554
25. Leng L, Zhang J, Xu J, Khan MK, Alghathbar K (2010) Dynamic weighted discrimination power analysis in DCT domain for face and palmprint recognition. In: 2010 International Conference on Information and Communication Technology Convergence (ICTC), pp 467–471
26. Leng L, Zhang J, Chen G, Khan MK, Alghathbar K (2011) Two-directional two-dimensional random projection and its variations for face and palmprint recognition. In: International Conference on Computational Science and Its Applications, LNCS, vol 6786, pp 458–470
27. Ling H, Soatto S, Ramanathan N, Jacobs DW (2007) A Study of Face Recognition as People Age. In: Proc of the IEEE 11Th International Conference on Computer Vision, ICCV 2007, pp 1–8
28. Lu X, Jain AK (2004) Ethnicity identification from face image. proc SPIE:114–123
29. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings of the IEEE international conference on automatic face and gesture recognition, FG 1998, pp 200–205
30. McGarry DP, Arndt CM, McCabe SA, D'Amato DP (2004) Effects of compression and individual variability on face recognition performance. SPIE Defense & Security Symposium, Biometric Technology for Human Identification, 5404-34 (pp. 362-372) International Society for Optics and Photonics
31. Meyers E, Wolf L (2008) Using biologically inspired features for face processing. Int J Comput Vis 76(1):93–104
32. Modi SK, Elliott SJ, Whetsone J, Hakil Kim (2007) Impact of age groups on fingerprint recognition performance. Proceedings of the 2007 IEEE Workshop on Automatic Identification Advanced Technologies:19–23
33. Ng W, Lindsay RC (1994) Cross-race facial recognition: Failure of the contact hypothesis. J Cross-Cultural Psychol 25:217–232
34. Ng CB, Tay YH, Goi BM (2012) Vision-based Human Gender Recognition: A Survey. Available online at arXiv:1204.1611. Accessed January 2015
35. O'Toole AJ, Phillips PJ, Jiang F, Ayyad J, Pénard N., Abdi H (2007) Face recognition algorithms surpass humans matching faces over changes in illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(9):1642–1646
36. O'Toole AJ, An X, Dunlop J, Natu V, Phillips PJ (2012) Comparing face recognition algorithms to humans on challenging tasks. ACM Transactions on Applied Perception 9(4):16
37. Phillips PJ, Wechsler H, Huang J, Rauss P (1998) The FERET database and evaluation procedure for Face-Recognition algorithms. Image Vis Comput J 16(5):295–306
38. Phillips PJ, Grother P, Micheals R, Blackburn DM, Tabassi E, Bone M (2003) Face recognition vendor test 2002. In: Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures,AMFG 2003
39. Phillips PJ, Flynn P, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVFR 2005
40. Phillips PJ, Scruggs W, O'Toole A, Flynn P, Bowyer K, Schott C, Sharpe M, FRVT 2006 and ICE 2006 large-scale experimental results (2010) IEEE Trans Pattern Anal Mach Intell 32 (5):831–846
41. Phillips PJ, Beveridge JR, Draper BA, Givens G, O'Toole AJ, Bolme DS, Dunlop J, Lui YM, Sahibzada H, Weimer S (2011) An introduction to the good, the bad, & the ugly face recognition challenge problem. In: Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011, pp 346–353
42. Rhodes MG, Anastasi JS (2012) The own-age bias in face recognition: a meta-analytic and theoretical review. Psychol Bull 138(1):146
43. Ricanek K, Tesafaye T (2006) MORPH: A longitudinal image database of normal adult age-progression. In: Proceedings of 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, pp 341–345
44. Riccio D, Tortora G, De Marsico M, Wechsler H (2012) EGA - Ethnicity,Gender and Age, a pre-annotated face database. In: Proceedings of 2012 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications, bioMS 2012, Salerno (Italy), September 14 2012, pp 38–45

45. Ross A, Jain AK (2004) Biometric sensor interoperability: A case study in fingerprints. Biometric Authentication. Lect Notes Comput Sci 3087:134–145
46. Shakhnarovich G, Viola PA (2002) Moghaddam,b., A unified learning framework for real time face detection and classification. In: Proceedings 5th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2002, pp 14–21
47. Spacek L The Essex database, http://cswww.essex.ac.uk/mv/allfaces/index.html , University of Essex
48. Sun Z, Bebis G, Yuan X, Louis SJ (2002) Genetic feature subset selection for gender classification: a comparison study. In: Proceedings of the WACV 2002, pp 165–170
49. Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp 1988–1996
50. Taigman Y, Yang M, Ranzato MA, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1701–1708
51. The FEI face database, Available at http://www.fei.edu.br/~cet/facedatabase.html, Accessed January 2015
52. Toderici G, O'Malley SM, Passalis G, Theoharis T, Kakadiaris IA (2010) Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques. Int J Comput Vis 89(2-3):382–391
53. Turk MA, Pentland AP (1991) Face recognition using eigenfaces. In: Proceedings of the International Conference on Pattern Recognition, ICPR 1991, pp 586–591
54. Vasuhi S, Vaidehi V (2009) Identification of human faces using orthogonal locality preserving projections. In: Proceedings of the 2009 International Conference on Signal Processing Systems, pp 718–722
55. Veropoulos K, Bebis G, Webster M (2005) Investigating the impact of face categorization on recognition performance. Proceedings of the ISVC, 2005:207–218
56. Zhao W, Chellappa R, Rosenfeld A, Phillips PJ (2003) Face recognition: a literature survey. ACM Comput Surv:399–458

Maria De Marsico is Associate Professor at Sapienza University of Rome, Department of Computes Science, where she has been Assistant Professor from 2001 to 2015. She got her Master degree in Computer science from University of Salerno in 1988. Her scientific interests focus on Human Computer Interaction and Image Processing. Regarding the first one, she is especially interested in problems related to accessibility for users with special needs, and advanced techniques for personalized distance learning. Regarding the second one, she works on biometric recognition, including face, iris, gate, and multimodal recognition. She is Associate Editor of Pattern Recognition Letters, and Area Editor of the IEEE Biometrics Compendium. She published about 100 scientific works in international journals, conferences, and book chapters. She has been member of many Technical program Committees and is referee for several top journals. She recently organized the 2015 edition of CHItaly, the biannual conference of the Italian chapter of ACM SIGCHI. She is member of IEEE, ACM and of the Italian Group of Italian Researcher in Pattern Recognition.



Michele Nappi was born in Naples, Italy, in 1965. He received the laurea degree (cum laude) in computer science from the University of Salerno, Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from I.I.A.S.S. "E.R. Caianiello", Vietri sul Mare, Salerno, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Padova, Italy. He is currently an Associate Professor of computer science at the University of Salerno. His research interests include Multibiometric Systems, Pattern Recognition, Image Processing, Compression and Indexing, Multimedia Databases, Human-Computer Interaction, VR/AR. He co-authored over 120 papers in international conference, peer review journals and book chapters in these fields (see http://www.informatik.uni-trier.de/ ley/pers/hd/n/Nappi:Michele.html). He also served as Guest Editor for several international journals and as Editor for International Books. In 2014 He was one of the founders of the spin off BS3 (Biometric System for Security and Safety). President of the Italian Chapter of the IEEE Biometrics Council (2015-2017), Member of IAPR and IEEE and team leader of the Biometric and Image Processing Lab (BIPLAB), Dr. Nappi received several international awards for scientific and research activities.



Daniel Riccio was born in Cambridge, U.K., in 1978. He received the Laurea degree (cum laude) and the Ph.D. degree in computer science from the University of Salerno, Salerno, Italy, in 2002 and 2006, respectively. He is currently an Associate Professor at the University of Naples, Federico II. His research interests include biometrics, image analysis and coding, and indexing. Prof. Riccio is a IEEE member since 2012. He is also a member of the Italian Group of Italian Researcher in Pattern Recognition since 2004.

Harry Wechsler (Fellow) received the PhD degree in computer science from the University of California, Irvine, in 1975, and serves now as Professor of Computer Science at George Mason University. His research covers intelligent systems, biometrics, image and signal processing, computer vision, data mining, statistical learning and model selection, pattern recognition, and security and privacy. The range of applications covers evidence-based and identity management, interoperability, face recognition, gait analysis, performance evaluation and error analysis, bioinformatics, change, malware, outlier, phishing, and spam detection, and video processing and surveillance. He organized and directed the NATO Advanced Study Institute (ASI) on "Face Recognition: From Theory to Applications" (Stirling, UK, 1997), and was the principal co – editor for its seminal proceedings published by Springer (1998). His book on Reliable Face Recognition Methods, which breaks new ground in applied modern pattern recognition and biometrics, was published by Springer in 2007. Dr. Wechsler directed at GMU the design and development of FERET, which has become the standard facial data base used for benchmark studies and experimentation. He was elected an IEEE Fellow in 1992 for "contributions to spatial/spectral image representations and neural networks and their theoretical integration and application to human and machine perception" and an IAPR (International Association of Pattern Recognition) Fellow in 1998. He was granted (together with his former doctoral students) five patents by US Patent Office (USPO) on fractal image compression using quad-q-learning (licensed in 2006), on feature based classification (for face recognition), on open set recognition and outlier detection using transduction, on adaptive and robust filters for face recognition, and on data stream change detector.