Demographics versus Biometric Automatic Interoperability

Maria De Marsico¹, Michele Nappi², Daniel Riccio², and Harry Wechsler³

¹ Department of Computer Science, Sapienza University of Rome, Italy
demarsico@di.uniroma1.it

² Biometric and Image Processing Lab, University of Salerno, Italy
{mnappi, driccio}@unisa.it

³ Department of Computer Science, George Mason University, Fairfax Virginia, USA
wechsler@cs.gmu.edu

Abstract. Humans are naturally experts in recognizing faces. Such skills are enforced through a mix of cultural and cognitive processes. This allows human vision system to be especially efficient and effective in processing faces in a familiar environment. Automatic recognition system are currently not able (will ever be?) to achieve similar performance, especially when cross-demographic features are involved (gender, ethnicity, and age). Recent studies suggest a significant decrease of the number of recognition errors by limiting the search space to faces with the same demographics. This can be obtained by preliminarily annotating faces with a demographic profile, or by using demographic features as soft biometrics to be determined as a support to actual recognition. Especially in the second case, a multi-demographics dataset is needed to appropriately train a recognition system, and/or to test its performance. In this paper we use EGA dataset to test how interoperability relationships between biometric and demographics can be exploited for better recognition, though avoiding human intervention to preventively select appropriate demographic parameters.

1 Introduction

Biometric recognition performance is affected by many factors, among them, protocols (“best practices”), performance evaluation and validation [2]. Protocols should enforce the proper and effective use of biometric systems, while performance evaluation and validation should take into account uncontrolled-settings and interoperability in order to determine the robustness and reliability of the biometric authentication decisions even when datasets used during system implementation are different from those actually used during normal system operation. A robust system should present a low sensitivity to biometric variability, including incomplete data (“occlusion”), corrupted data (“disguise”), as well as pose, illumination and expression (PIE) variations. More and more studies recently focus on further variability elements, namely demographics (gender, ethnicity, and age), since they revealed a strong influence on recognition performance, if the datasets used for implementation do not present the same range of variations as the actual population to be processed. This also affects

reliability, i.e. the consistency and stability of the predictions made on authentication. Both robustness and reliability are therefore influenced by variability of information due to different factors. Uncontrolled settings are especially widespread in real life scenarios, e.g., mass screening, surveillance using smart camera networks, and tagging within social networks. Performance for such settings is definitely significantly lower than the performance for tightly controlled biometrics evaluations, particularly in large scale authentication scenarios. This paper evaluates different operational biometric scenarios for the purpose of interoperability in different demographic settings, with demographic composition and quality of biometric data the varying factor. In general, interoperability is the ability of diverse systems and/or organizations to work together (inter-operate) through well-defined tasks and protocols. In the biometric context, it refers to suites of sensors, data flows, subjects' condition, and algorithms. Especially when demographic differences are involved, we can consider interoperability as the link between biometrics and forensics, within the context of distributed data collection and associated federated identity management systems [2]. Biometric interoperability is addressed here in terms of diversity of population enrolled.

Human observers of a face can easily make subtle judgments for gender, identity, age and expression. These categorizations reasonably exploit different visual information, though it is not possible, in general, to isolate the specific information they use. Many attempts to investigate such issue mainly pertain to vision research. Human skills are enforced through a mix of cultural and cognitive processes, evolution (phylogeny) within a specific ethnic group as well as development (ontogeny) of a person within a mixed population, including different ages and genders. These processes allow the human vision system to be especially efficient and effective in processing faces in a familiar environment. The implicit, unique combination of such processes may also explain some controversial phenomena known to undermine human face recognition. The most known one is the other-race effect, i.e. the poor ability to recognize a subject from a different race. Though often ascribed to the so-called "contact hypothesis" [1], which refers to the single individual experience, it cannot be completely explained by it [7]. On the other hand, the concept of "familiarity" is often the best explanation for excellent human recognition performance. The investigation of cognitive processes regarding cross-demographic features affecting human face recognition ability, and their relevance to automatic systems, is expected to significantly improve face recognition performance. Even for computer systems specific problems arise when cross-demographic features are involved. This phenomenon was seldom systematically observed before the 2006 NIST Face Recognition Vendor Test (FRVT) [9]. The simple reason was that biometric systems implemented in a specific country, e.g. Asian, would usually rely on a home-made dataset of subjects of the same ethnicity. Therefore the cross-demographics interoperability was hardly a problem. The main obstacle for interoperability is that biometric face templates are mostly obtained by extracting the same features regardless the population at hand. It is, however, clear that different categorizations (age, race, ethnicity gender, etc.) should rely on different features. A comprehensive study of the influence of demographic features on recognition often suffers from the lack of large datasets with a wide, uniform variation of such features and possibly pre-annotated data [11]. If these features are significant

categorization criteria, one can conjecture that about the same number of subjects from each group should be represented in the dataset used to train biometric systems, to assure robustness and reliability even in contexts different from the implementation one. The work by Klare *et al.* [4] is the most recent attempt to systematically investigate the role of cross-demographics in automatic face recognition.

Regarding interoperability achieved through the use of different datasets, the experiments in [2] demonstrate a significant dependence between recognition performance and the use of different datasets for training and testing. Experiments in [4] show that such dependence is even heavier if demographics differences are also involved. Along the same line, we present experiments performed using EGA dataset (Ethnicity, Gender and Age face database) [11] to further investigate strategies to avoid human intervention in pre-selection of subsets related to specific demographics.

2 The EGA Face Database

One of the most crucial limits for the growing number of works in the area of face recognition, also including multimodal recognition, face age/gender/ethnicity estimation, and face re-identification, is the lack of a face dataset sufficiently large and variegated. Most studies still refer to FERET dataset [8], which to the best of our knowledge, is still the one with the largest number of subjects including good diversification of features related to ethnicity, gender, illumination, and pose. FERET, however, is lacking in both the total number of subjects, and in their overall demographic heterogeneity. The latter is found in the dataset used in [4], namely the set of one million mug shot face images from the Pinellas County Sheriff's Office (PCSO). However, these lack a sufficient amount of Hispanic subjects, as well as younger (<18) or elder (>50) people. EGA dataset (Ethnicity, Gender and Age face database) [11] which is used here, has the advantage to be flexibly updateable and extendable in time to increase the amount of significant variations. In other words, the set of its images is not fixed once and for all, but can be extended according to the needs of the experiments at hand. EGA integrates into a single dataset more sets of face images from different databases which are available nowadays, and are quite different depending on the place and the context where they were gathered. Templates are further organized according to features such as ethnicity, gender and age.

EGA is gathered through a set of links to files that have been processed in advance by appropriate scripts. Each user can ask and obtain on her own the original datasets with the images needed to build the needed version of EGA, whether they are free or require some form of registration/payment. The scripts reorganize and rename the requested images, according to the structure that was devised for EGA. It is not only possible to easily build the dataset, but even to expand it, as new datasets become available and after they are annotated. Thanks to the annotation of single images, it is also possible at any moment, to discard specific samples. In the present composition, distortions such as illumination, pose, expression and occlusions are minimized, since representative datasets are already available, while demographic aspects are appropriately emphasized. Figure 1 shows the conceptual organization and the current composition of EGA.

Fig. 1. Logical organization and composition of EGA

EGA v1.0 gathers images from six different datasets.

- CASIA-Face V5 [12]: 2,500 images of 500 subjects, captured in a single session with resolution 640×480, 16bit color. Faces possibly present distance, pose, illumination, and expression variations, including wearing eyeglasses. Subjects are mostly young and of Eastern ethnicity.
- FEI [13]: 2,800 images of 200 subjects (100 male and 100 female), with 14 color images each of 640×480 resolution. Faces have been acquired on a white background and belong to subjects from 19 to 40 years old, mostly of Latin ethnicity.
- FERET [8]: 14,126 images of 1,199 subjects of different ethnicities, genders and age, acquired in 15 sessions between August 1993 and July 1996. Resolution is 256×384 for 8bit greyscale. Different face sets (fa, fb, dup I, dup II) differ in pose and acquisition period, and present slight variations in illumination and expression.
- FRGC [9]: 50,000 images from 4,003 subject sessions (for each session four images in controlled conditions, two images in non-controlled conditions, and a 3D model). Images are divided into training and validation partitions. Resolution is 1704×2272 for 24bit color. Most subjects are Caucasians, and are mainly concentrated in a same age range (young/adult). Gender is quite uniformly represented.
- JAFFE [5]: 2130 images of 10 subjects. It was mainly gathered for facial expression analysis. Image resolution 256×256 for 8bit greyscale. Subjects are all female and Japanese of uniform age.

- Indian Face Database [3]: the dataset contains 40 subjects in frontal pose with eleven different looking directions for each individual, plus some additional image when available. Images are captured on a uniform background, and further present four expression variations. The dataset includes one folder for males and one for females. Image resolution is 640×480 pixels, with 256 grey levels. All subjects are of Indian ethnicity, with an adequate distribution for gender, but not for age.

3 Proposed Approach

Experiments by Klare *et al.* in [4] highlighted, among others, two relevant aspects:

- Training a recognition system on a data set of heterogeneous images, does not reduce limitations of its performance with respect to “hardest” ethnic groups, i.e. those which are more difficult to recognize.
- The performance of a recognition system for a specific ethnicity is improved when it is trained on that ethnic group.

Klare *et al.* suggest, therefore, that the priori selection of the ethnicity (performed by a human operator) of an input image can produce an increased accuracy of the recognition system. However, notwithstanding the precise and accurate analysis of all the assumptions, this hypothesis is not extensively verified in the referenced work. Moreover, the suggested exploitation of the presented findings requires human intervention in the preliminary selection of the ethnicity of each face. The authors assume that “in most operational scenarios, particularly those dealing with forensics and law enforcement, the use of face recognition is not being done in a fully automated, “lights out” mode. Instead, an operator is usually interacting with a face recognition system, performing a one-to-one verification task, or exploring the gallery to group together candidates for further exploitation” [4]. However, in further equally significant scenarios, such as video-surveillance or access control, continuous human interaction might not be feasible. We want to explore viable strategies for this second hypothesis, which allow to avoid the process of human demographics pre-selection. We still rely on multiple classifiers, each trained on a specific demographic feature. However, and this is the core of our work, we rule out any intervention during normal system operations, except for particularly hard cases where an appropriate alert system may signal the need for an operator’s check. There are two different ways to achieve this, both relying on a pool of classifiers, each trained for a specific demographic feature:

1. A Priori Demographics Selection (APrDS): thanks to automatic recognition of relevant demographic features (ethnicity, gender, age, or a combination of them), each probe image is submitted to the corresponding classifier; this requires a further training for a first identification of the involved soft biometrics classes, and a further preliminary step during each recognition operation.
2. A Posteriori Demographics Selection (APoDS): it is assumed that the demographic features of the subject are not known a priori, therefore the probe image is inputted to all the classifiers; to complete the recognition process, it is

necessary to adopt a criterion for the selection of the global best answer; this is the real novelty of our proposal, since we also rule out any kind of automatic demographics preprocessing.

In the experiments discussed here, we simulated both kinds of system. In particular, we chose ethnicity as the first test-bed, since this feature usually presents the highest number of different significant classes, and therefore is prone to the highest degree of confusion. In the specific case, since demographics (soft biometrics) recognition is not our final goal, the APrDS system relied on using EGA as dataset. In particular, the selection of ethnicity of a probe image was simulated by exploiting the metadata provided by the nomenclature of the dataset. On the other hand, a APoDS system requires more attention. This kind of system is structured in such a way that the same probe image is supplied to the different classifiers, which each perform recognition in relation to their own gallery (consisting of images of a specific ethnic group) and provide as output a similarity score s and a reliability margin φ ; in this work we adopted as φ the function defined as (more details are available in [6]):

$$\varphi(p) = 1 - |N_b|/|G| \quad \text{with} \quad N_b = \{g_{i_k} \in G \mid F(d(p, g_{i_k})) < 2 \cdot F(d(p, g_{i_1}))\} \quad (1)$$

where $|G|$ is the gallery cardinality, d is a distance measure, g_{i_k} is a gallery element (template), p is a probe template, g_{i_1} is the first element returned in the list corresponding to the identification of p , and F is a score normalization function (we used here the Quasi-Linear Sigmoidal (QLS) [6]). The architecture is sketched in Figure 2.

Fig. 2. The architecture of APoDS system

Following EGA composition, we have five classifiers based on ethnicity: Afro-Americans, Asians, Caucasians, Indians and Latinos. The system includes a Selector module, which collects the scores and reliability margins produced by the involved classifiers and selects the best response (classifier) as follows:

$$cl = \arg \max_j \{s_j * \varphi_j\}, j = 1, 2, \dots, 5 \quad (2)$$

In other terms, the Selector selects the response from the classifier which achieves the maximum product of the pair (s_j, q_j) . It is worth noticing that our APoDS approach addresses also the mixed ethnicities, since taking the most reliable response from the classifiers implies to exploit the dominant ethnic features for each probe subject.

4 The Experiments

The experiments were carried out on images from EGA. They were focused on ethnicity, leaving to future work the verification of the same issues regarding gender and age. Ethnicity was preferred, as it provides a greater degree of variation (five possible classes in EGA) than the others (two for gender and three for age). We used widely available classifiers (LDA and OLPP), in order to obtain baseline and well repeatable results about the investigated issues. The general adopted protocol was the following. For each subject, 5 images were considered (1, 2, 3, 4, 5). For each of three experimental rounds three images were used for training, 1 for the gallery, and 1 as the probe. The reported Recognition Rate (RR) represent the average over the three rounds of subject identification:

- 1 round: 1,2,3 (training), 4 gallery, 5 probe
- 2 round: 2,3,4 (training), 5 gallery, 1 probe
- 3 round: 3,4,5 (training), 1 gallery, 2 probe

The first part of the experiments aimed at ensuring that the behavior of the chosen classifier was consistent with the outcomes by Klare *et al.* in [4]. The first experiment aimed at assessing how and to what extent the training set used for a given classifier affects its performance, if it is applied in a different context. In this case, the LDA classifier was trained on the five different ethnicities: African-Americans (AA), Asians (AS), Caucasians (CC), Indians (IN) and Latinos (LT). The experiment was repeated five times, once for each ethnic group. Each time, the probe and gallery images were made up of only members of the ethnic group under consideration. On the other hand, the same classifier was trained on a single ethnic group, either the same or a different one with respect to testing. After it was applied to the pair probe / gallery, RR was computed. Table 1 shows RR of the LDA classifier with single-ethnicity training sets, when applied to probe / gallery sets belonging to a single ethnicity.

Table 1. RR of LDA classifier trained (rows) and tested on single ethnicities

LDA	AA	AS	CC	IN	LT
AA	0.62	0.56	0.66	0.71	0.58
AS	0.74	0.74	0.73	0.77	0.74
CC	0.70	0.70	0.71	0.64	0.65
IN	0.70	0.72	0.72	0.69	0.73
LT	0.68	0.74	0.72	0.66	0.69

Intuitively, we would expected that a classifier trained on images belonging to a specific ethnic group, gives the best result of the probe / gallery images belonging to

the same ethnic group. The results in Table 1 seem to refute this hypothesis. We initially assumed that this phenomenon could depend on the different number of subjects available for each ethnic group. Therefore we chose the ethnicity with the minimum number of subjects, and limited the set of images to this value for all ethnic groups, so as to ensure an even distribution for the different ethnic groups. However, the behavior of the LDA classifier was found to be entirely consistent with that seen before. The second hypothesis focused on a further factor, extremely relevant to take into consideration, which is the sensitivity of the classifier to any distortion in the images. In fact, it is known from the literature that the LDA, although more robust to distortions than other dimensionality reduction techniques, is not particularly accurate, when such distortions are particularly relevant. The images in the EGA database were selected in such a way as to be all in frontal pose and nearly neutral expression. However, coming from different datasets, they show different illuminations (an important aspect in the LDA) and different resolutions. In order to verify whether the inconsistent behavior was actually determined by the lack of robustness of the classifier, the same experiment was performed with a different, more robust classifier, i.e. OLPP. Table 2 shows now a behavior which is fully consistent with the expected outcomes: a classifier which is relatively robust to other sources of distortion, once trained with images from a certain ethnicity, achieves the best RR values with probe/gallery sets which only contain images from the same ethnicity.

Table 2. RR of OLPP classifier trained (rows) and tested on single ethnicities

OLPP	AA	AS	CC	IN	LT
AA	0.67	0.62	0.64	0.65	0.62
AS	0.69	0.73	0.72	0.71	0.71
CC	0.67	0.69	0.73	0.68	0.69
IN	0.72	0.71	0.70	0.73	0.71
LT	0.73	0.73	0.76	0.75	0.77

Given these results, it is particularly interesting to verify the behavior of the same classifier, OLPP from now on, when it is trained on a single ethnicity but tested on probe/gallery sets with images from all ethnic groups (all five in this case). Table 3 shows results of this experiment.

Table 3. RR of OLPP classifier trained on single ethnicities and tested on probe/galleries with five ethnicities

OLPP	AA	AS	CC	IN	LT
	0.66	0.69	0.70	0.68	0.68

In all cases in Table 3, we observe that RR is lower to that achieved by the same classifier when it is trained and tested on a same specific ethnicity.

In the opposite situation, the classifier was trained on images from all ethnic groups, and tested on a single one. Table 4 shows the results.

Table 4. RR of OLPP classifier trained on five ethnicities and tested on probe/galleries with single ethnicities

OLPP	AA	AS	CC	IN	LT
	0.64	0.72	0.70	0.73	0.75

It is clear that the classifier in question provided the best performance in terms of RR when working independently on individual ethnic groups.

Generally, in a system functioning in a real world application, it is much more likely that the classifier is trained on a set of images that is less representative and varied than that on which it will then be applied at operating time. This is because companies and laboratories have limited datasets. Furthermore, most existing systems do not make a preliminary judgment about the ethnicity of the acquired probe image, therefore the case more likely to be considered for comparison is one in which the classifier is trained on all ethnic groups and is then tested on all ethnicities. By performing such a test, in this specific case, we obtained a RR of 0.68, which is well below the performance achieved by the same classifier, when applied to the individual ethnic groups. The results suggest, as in [4], that if it was possible to identify a priori the ethnicity of the subject and submit the image to the classifier trained for that specific ethnicity, the accuracy would be greater. By performing this experiment we observed that the RR rises to 0.73. We remind that the correct determination of the ethnicity was simulated in this experiment of APrDS, using the nomenclature of files in EGA.

Finally, we tested the accuracy of a APoDS simulation. We assumed an unknown ethnicity of the input subject. We separately submitted the image to each classifier trained on individual ethnic groups, and verified if it was possible, from time to time, to retrieve the response (possibly correct) of the classifier corresponding to that ethnicity. The RR provided by the APoDS simulation is 0.67, which is comparable with the performance of the system trained and tested on all ethnic groups (0.68). However, this value is still less than what one would get with a APrDS system (RR 0.73). This suggests that, although it is possible to improve the performance of an identification system by considering the different ethnic groups (RR 0.73), it is a complex problem to be able to do this by selecting a posteriori the response of the most suitable classifier, or in other words, inferring ethnicity a posteriori.

5 Conclusions

An increasing number of papers addresses the problem of interoperability of trainable classifiers with respect to demographic variations. However, systematic studies of this phenomenon are still few. This work has the goal to evaluate the weight of ethnicity on the training / testing of a classifier, and to verify the hypothesis that the combination / selection of classifiers, each trained on a separate ethnic group, can produce an increased accuracy in terms of RR. The results obtained show that, once the ethnicity of the subject is known a priori, it is possible to actually increase the RR by about 7%. In contrast, a system that selects the correct answer downstream of the recognition process (a posteriori estimate of ethnicity) is able to achieve performance at least

comparable to those of the system trained / tested on all ethnic groups, but still further improvable. A possible topic for future work will be the identification of effective criteria to apply implemented by the module Selector to this aim. Moreover, we will test more demographics and their possibly combined influence on recognition.

References

1. Chiroro, P., Valentine, T.: An investigation of the contact hypothesis of the own-race bias in face recognition. *Quarterly J. Experimental Psychol. Human Experimental Psychol.* 48A, 879–894 (1995)
2. El Khiyari, H., De Marsico, M., Abate, A.F., Wechsler, H.: 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, pp. 1–8 (2012)
3. Jain, V., Mukherjee, A.: The Indian Face Database (2002), <http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/>
4. Klare, B.F., Burge, M.J., Klontz, J.C., Vorder Bruegge, R.W., Jain, A.K.: Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security* 7(6), 1789–1801 (2012)
5. Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding Facial Expressions with Gabor Wavelets. In: *Proceeding of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205 (1998)
6. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: NABS: Novel Approaches for Biometric Systems. *IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews* 41(4), 481–493 (2011)
7. Ng, W., Lindsay, R.C.: Cross-race facial recognition: Failure of the contact hypothesis. *J. Cross-Cultural Psychol.* 25, 217–232 (1994)
8. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. *Image and Vision Computing J.* 16(5), 295–306 (1998)
9. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA (2005)
10. Phillips, P., Scruggs, W., O'Toole, A., Flynn, P., Bowyer, K., Schott, C., Sharpe, M.: FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(5), 831–846 (2010)
11. Riccio, D., Tortora, G., De Marsico, M., Wechsler, H.: 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, pp. 38–45 (2012)
12. CASIA-FaceV5, <http://biometrics.idealtest.org/>
13. The FEI face database (2012), <http://www.fei.edu.br/~cet/facedatabase.html>