On the Use of Discriminative Cohort Score Normalization for Unconstrained Face Recognition

Massimo Tistarelli, *Senior Member, IEEE*, Yunlian Sun, and Norman Poh, *Member, IEEE*

***Abstract*— Facial imaging has been largely addressed for automatic personal identification, in a variety of different environments. However, automatic face recognition becomes very challenging whenever the acquisition conditions are unconstrained. In this paper, a picture-specific cohort normalization approach, based on polynomial regression, is proposed to enhance the robustness of face matching under challenging conditions. A careful analysis is presented to better understand the actual discriminative power of a given cohort set. In particular, it is shown that the cohort polynomial regression alone conveys some discriminative information on the matching face pair, which is just marginally worse than the raw matching score. The influence of the cohort set size in the matching accuracy is also investigated. Further, tests performed on the Face Recognition Grand Challenge ver 2 database and the labeled faces in the wild database allowed to determine the relation between the quality of the cohort samples and cohort normalization performance. Experimental results obtained from the LFW data set demonstrate the effectiveness of the proposed approach to improve the recognition accuracy in unconstrained face acquisition scenarios.**

***Index Terms*— Biometric verification, face recognition, cohort score normalization.**

I. Introduction

PEOPLE can naturally recognize others from their face appearance. The human visual system is capable of performing this task very quickly and almost effortlessly. The current automatic face recognition systems can also perform this task reasonably well in a number of practical applications, whenever the face image capture process is controlled or constrained [1]. However, several challenges still need to be addressed for unconstrained face recognition. In this setting, the face images of the same person appear very differently due to variability in the acquisition environment (e.g., under different illumination conditions), facial expression, the interaction with the face acquisition device (different head poses) and the alteration of the face traits due to either natural factors [2] or plastic surgery [3]. In order to deal with these variations effectively, earlier efforts have been mainly devoted to recognize faces collected in controlled lab environments. A number of face databases have been assembled to understand the effects of variability due to head poses, lighting conditions, expressions and occlusions. With these databases, many face recognition algorithms have been developed. According to the type of features used, the existing algorithms can be broadly classified into holistic and local methods. Subspace and manifold learning methods are among the holistic methods [4], [5]. Local methods include the widely used Gabor wavelets [6], local binary patterns (LBP) [7] and scale-invariant feature transform (SIFT) [8]. With these developed algorithms, face recognition in controlled conditions has achieved impressive improvement in performance over the years [9].

Face recognition in unconstrained scenarios has two relevant applications: surveillance and semantic web search. With the popularity and increasing number of video surveillance cameras installed in public places, face recognition is no doubt an important instrument in the fight against crime. At the same time, social networks such as Picasa and Facebook have generated an unprecedented volume of photos and videos. Automatic face recognition will play an increasingly important role to improve speed and efficiency in retrieving contents. For instance, photo tagging is a convenient feature that allows the end user to quickly retrieve photos of friends and family. Face images stored in social networks or found in surveillance videos are often captured in uncontrolled environments. These applications call for more robust automatic face recognition algorithms [10].

Face recognition can refer to a number of different tasks including, but not limited to: face identification, face verification and face pair matching [11]. Given a probe face image (or video), *face identification* aims to establish the identity of the individual from a gallery set of users. In *face verification*, the goal is to decide whether the identity of a submitted (query) face image (or video) is the same as the one claimed by the user. Similarly, *face pair matching* aims to determine whether two pictures represent the same individual or not. While for face identification and verification some statistical information on the user distributions as well as more images can be collected and available, in the case of face pair matching, the only available information is the photometric data contained in the two pictures. The lack of additional information makes face pair matching particularly difficult.

Manuscript received March 24, 2014; revised July 4, 2014; accepted October 3, 2014. Date of publication October 8, 2014; date of current version November 10, 2014. This work was supported in part by the European Commission COST Action under Grant IC1106 and in part by the Biometrics Evaluation and Testing Project through the European Union within the FP7 Programme under Grant 284989. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Edwin R. Hancock.

M. Tistarelli and Y. Sun are with the Department of Sciences and Information Technology, University of Sassari, Sassari 07100, Italy (e-mail: tista@uniss.it; yunlian.sun2@unibo.it).

N. Poh is with the Department of Computing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: n.poh@surrey.ac.uk).

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

Digital Object Identifier 10.1109/TIFS.2014.2362007

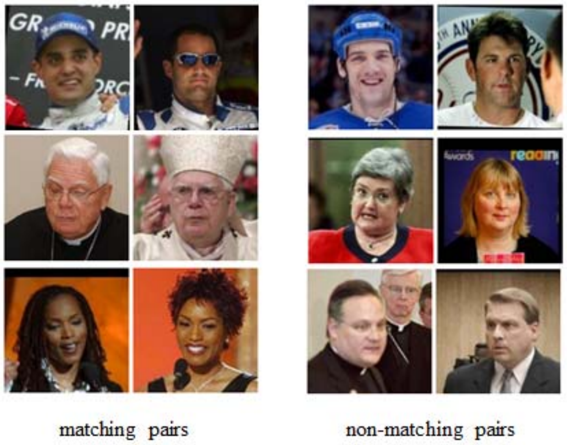

Fig. 1. Examples of matching and non-matching pairs from the LFW database.

The Labeled Faces in the Wild (LFW) database [12] is a relatively new benchmark for evaluating algorithms for unconstrained face pair matching. Faces in this database are collected from news articles in the web embedding a large and unpredictable variability. Fig. 1 shows some matching (two images are from the same person) and non-matching (two pictures are of different subjects) pairs from this database.

In recent years, cohort samples (identities of cohort samples are different from those of samples being compared) have been extensively used to improve the recognition performance of a biometric expert [13], [14]. These approaches have often been referred to as *cohort score normalization*. In this paper, we exploit the usefulness of this approach for matching face image pairs, captured under unconstrained conditions. In particular, it is worth showing whether post-processing the raw matching scores using cohort normalization can achieve performance which is comparable to the state-of-the-art algorithms for unconstrained face pair matching. In addition, to achieve a better understanding of cohort behavior, an extensive experimental exploration on both the LFW database and the Face Recognition Grand Challenge (FRGC) ver2.0 database [15] will be presented.

This paper encompasses a preliminary work reported in [16] and yet provides a better understanding of discriminative cohort behavior. The main contributions in this paper include:

1) *Picture-Specific Cohort Normalization for Face Pair Matching:* For each picture in the image pair, an exclusive cohort score list is composed. Some discriminative information is then extracted from the two cohort score lists for score normalization.
2) *Comparison With the State-of-the-Art Methods:* The proposed system is compared against the state-of-the-art algorithms using the LFW database.
3) *Better Understanding the Behavior of Cohort Normalization:* In particular, four important issues are addressed:
   a) How much discriminative information is contained in the cohort samples alone? This discriminative information can be empirically quantified in terms of Equal Error Rate (EER) [17].
   b) How do the choice and the size of the cohort set affect the normalization performance?
   c) What is the result of employing cohort samples of different quality?
   d) Should a cohort set contain as many as possible subjects (each subject with the fewest possible samples) or as few as possible subjects (each subject with the utmost possible number of samples)?

II. Related Work

In this section, a concise literature review on unconstrained face recognition and cohort score normalization is reported.

A. Unconstrained Face Recognition

Since its release in 2008, the LFW database has received a considerable attention for benchmarking face recognition algorithms. Several algorithms were also developed, specifically for handling large amounts of real-world face images [18]. Among these algorithms, quite a few focus on designing powerful facial descriptors, either handcrafted or data-driven. Some examples are the patch-based LBP codes [19], the learning-based (LE) descriptor [20], the discriminant face descriptor (DFD) [21], the local quantized patterns (LQP) [22] and the local higher-order statistics (LHS) [23]. Other methods, instead of devising an elaborated representation of the face, aim to learn an appropriate similarity measure to better compare pairs of unconstrained samples. These metric learning-based techniques have shown a great potential. Logistic Discriminant Metric Learning (LDML) [24], Cosine Similarity Metric Learning (CSML) [25], Pairwise-constrained Multiple Metric Learning (PMML) [26] and Similarity Metric Learning over the intra-personal Subspace (Sub-SML) [27], are some representative algorithms.

It is worth noting that in unconstrained face pair matching, there is no additional information to better drive the matching. The only available data is the photometric information embedded in the image pair. To compensate for the lack of information, many recent approaches first assemble an independent background face database to extract some useful information to help the matching. Generally, the background face database does not contain pictures of the subjects appearing in the two images being compared. From a set of background samples, Wolf. et al defined several similarity functions to learn a discriminative model exclusive to the pair of images being compared: One-Shot Similarity (OSS), Two-Shot Similarity (TSS) and ranking similarity [19]. In [28], an additional identity data set was employed for building a set of either attribute or simile classifiers. For comparing two faces under significantly different settings, Yin et al. proposed to “associate” one input face with alike identities from an extra generic identity data set [29]. Liao et al. [30] proposed an alignment-free sparse representation approach for partial face recognition. In this approach, the gallery descriptors were extracted from a set of background face images together with one of the two images being compared. In [31], an independent training set was organized to build a Gaussian Mixture Model (GMM) [32] from the spatial-appearance features.

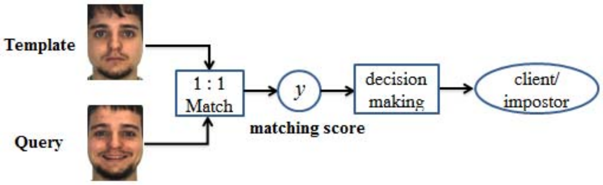

Fig. 2. A sample face verification system.

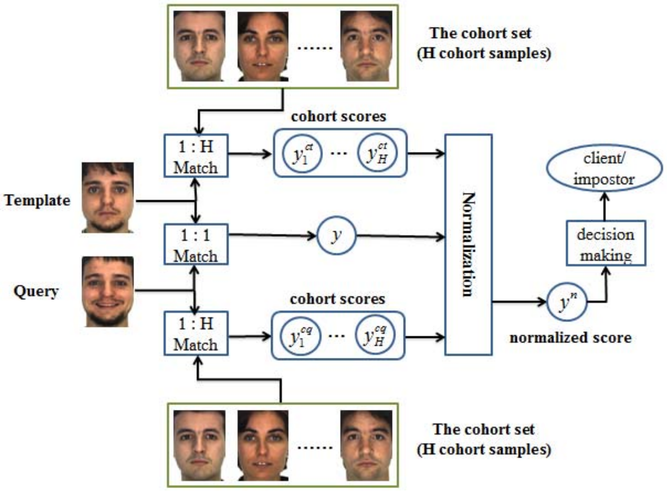

Fig. 3. A sample face verification system augmented with cohort score normalization.

B. Cohort Score Normalization

In a typical biometric verification system, the decision on the identity of the biometric sample is directly based on the matching score between the query sample and the claimed template model. Due to various sources of noise in the data capturing process, the biometric samples are often degraded, making the straightforward usage of the raw matching score unreliable. Therefore, post-processing the raw matching score, or score normalization [33], [34], becomes an essential stage. However, in many practical applications, only very few samples (or just one) are available for each subject, making it difficult to estimate the statistics of the client and impostor classes. Cohort-based score normalization is a technique used for mapping the raw matching score to a domain where the corrupting effect, caused by the large variability on the data, is reduced. Some information from a set of cohort samples, i.e., non-matching samples/impostors of the claimed identity, is required.

Cohort models have been proposed to model language processing and lexical retrieval [35]. For biometric applications, this technique was initially proposed for speaker recognition [13], [14]. In the literature, the term "background model" was also used to indicate the same concept [14]. This technique has been successfully applied to fingerprint verification [36], face verification [37], multi-biometrics [38] and under-sampled face identification [39]. Figures 2 and 3 show a conventional face verification system and the same system augmented with cohort normalization (face images are from the AR database [40]). A set of cohort scores is obtained by matching either or both of the two face images being compared with the cohort samples. Score normalization is performed by either estimating the score distribution parameters from the computed cohort scores, or extracting auxiliary information from the sorted cohort scores.

In the literature, many cohort-based score normalization approaches have been proposed. Zero-normalization (Z-norm) [41] and Test-normalization (T-norm) [41] are the two most common algorithms adopted in practical biometric applications. As both techniques assume a Gaussian distribution for each subject class, the first and second order moments of the cohort scores are computed for scaling the distributions. However, the cohort scores used in the Z-norm are the matching scores between the template and the cohort samples, while those applied in the T-norm are the matching scores between the query and the cohort samples.

In addition to the standard estimation and scaling involved in the Z-norm and T-norm, the methods proposed in [37], [38], and [42] attempt to further exploit the patterns of the sorted cohort scores. Among these approaches, the polynomial regression-based cohort normalization [37] has achieved promising results in some biometric applications. This technique drew its motivation from the observation that cohort samples, sorted by their similarity to the claimed target model, produced a discriminative pattern. In [37] polynomial regression was proposed to extract this discriminative information. In this approach, the matching scores of the cohort samples are computed against both the query sample and each enrolled template to determine a user-specific cohort rank ordering.

III. Picture-Specific Cohort Score Normalization

In principle, a subject-specific face representation allows to maximize the discrimination capability for each individual [44]. This approach requires to develop a computational model which embodies peculiar information for each subject.[1] By tailoring the analysis to each user, any identity claim is adapted to the user, or more precisely to the model associated with the user. This is accomplished by applying a polynomial regression-based cohort normalization to face pair matching scores. This process can normalise the variations in the score distribution due to the appearance of the two faces in a given pair of images. Fig. 4 schematically depicts the picture-specific cohort normalization process.

The proposed computational model also agrees with some psychophysical findings [43]. These suggest that the human visual system encompasses a model formation process for objects, including faces, where several continuous comparisons with other objects, or faces, are made i.e., by performing a repeated, comparative analysis, or pair-wise matching.

A. Picture-Specific Cohort Ordering

In a face verification system, the distribution of the cohort scores, obtained by matching a number of impostor and

[1] It is worth noting that the hypothesis space, made upon the score distributions of the probe and gallery samples belonging to the same person (match; genuine user/client; positive class) or a different one (non-match; impostor; negative class), can be very different from one user model to another.

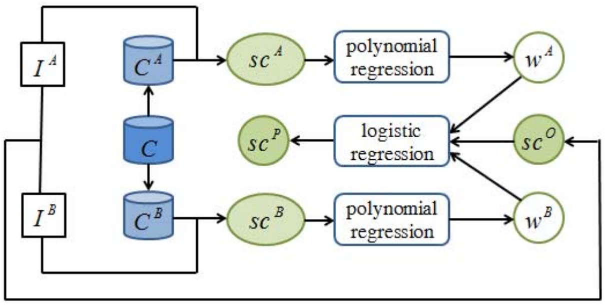

Fig. 4. Schematic process of picture-specific cohort score normalization for face pair matching.

genuine query samples with cohort samples, sorted with respect to their similarity to the claimed template, exhibits a discriminative pattern [37]. Therefore, it is reasonable to assume that in the face pair matching scenario, sorted cohort scores of matching pairs and non-matching pairs imply similar discriminative patterns. This assumption will be further verified in the experimental section.

Let (I^A, I^B) denote an image pair to be compared and sc^O be the corresponding raw matching score. Given an additional fixed cohort set $C = \{c_1, \ldots, c_h, \ldots, c_H\}$, composed of H cohort samples, an exclusive cohort list for each of I^A and I^B, named C^A and C^B, is composed. Both C^A and C^B are sorted variants of C, the only difference among the three sets lies in the rank order of the cohort samples.

A set of cohort scores are computed by matching each picture of the pair and all the cohort samples in C. C^A is obtained by sorting all cohort samples with respect to their closeness to I^B. In other words, $\{c_1^A, \ldots, c_h^A, \ldots, c_H^A\}$ are the H sorted cohort samples in C^A, where c_1^A is the most similar cohort sample to I^B, while c_H^A is the most dissimilar one. In the same way the cohort list $\{c_1^B, \ldots, c_h^B, \ldots, c_H^B\}$ for picture I^B is assembled, where c_1^B is the most similar cohort sample to I^A. Two picture-specific cohort score lists $sc^A = \{sc_1^A, \ldots, sc_h^A, \ldots, sc_H^A\}$ and $sc^B = \{sc_1^B, \ldots, sc_h^B, \ldots, sc_H^B\}$ are then obtained. The H scores in sc^A are all the matching scores between I^A and each cohort sample in C^A, where sc_i^A is the matching score between image I^A and the cohort sample c_i^A.

In summary, two picture-specific cohort score lists are generated using cohort samples sorted by the respective matching scores. This is different from the approach proposed in [37], where only one user-specific cohort score list is generated by cohort samples sorted according to their closeness to the claimed template. Given the two cohort score lists sc^A and sc^B, the embedded discriminative patterns are extracted by means of a polynomial regression.

B. Extraction of Discriminative Patterns Using Polynomial Regression

The sorted cohort scores are first considered as discrete points on a function of rank orders. More specifically, given the two picture-specific cohort score lists sc^A and sc^B, the following two rank orders functions are obtained:

$$sc_h^A = f^A(h) \quad (1)$$
$$sc_h^B = f^B(h) \quad (2)$$

where $h = 1, 2, \ldots, H$. Polynomial regression is applied to approximate the two functions as follows:

$$f^A(h) \approx w_n^A h^n + w_{n-1}^A h^{n-1} + \ldots + w_1^A h + w_0^A \quad (3)$$
$$f^B(h) \approx w_n^B h^n + w_{n-1}^B h^{n-1} + \ldots + w_1^B h + w_0^B \quad (4)$$

where $w^A = [w_0^A, w_1^A, \ldots, w_n^A]$ and $w^B = [w_0^B, w_1^B, \ldots, w_n^B]$ are the two approximated polynomial coefficient vectors. The cohort scores in sc^A and sc^B can be approximated by the $n+1$ coefficients in w^A and w^B respectively. Therefore, the coefficients in w^A and w^B can be used to represent the discriminative patterns embedded in the sorted cohort scores.

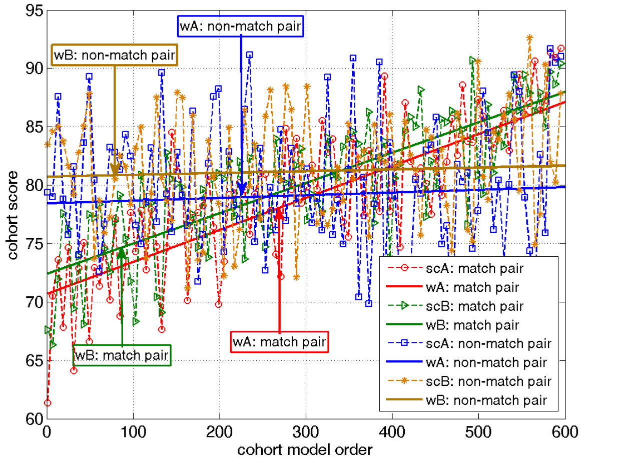

Fig. 5. Cohort score profiles and the corresponding fitting lines for a matching pair and a non-matching pair, computed from the LFW database.

In order to demonstrate the effectiveness of w^A and w^B to distinguish matching pairs from non-matching pairs, the cohort score profiles (i.e., sc^A and sc^B) of a matching pair and a non-matching pair, computed from the LFW database, as well as the fitted curves (i.e., w^A and w^B), are presented in Fig. 5. In this example the polynomial degree is simply set to 1, and a linear function is fitted through the cohort score profiles. As it can be noticed, the profiles of the cohort matching scores plotted against the cohort rank order are very noisy. Therefore, it is difficult to extract any discriminative information directly comparing the two cohort profiles. As shown in Fig. 5, by applying the polynomial regression the noise is largely suppressed, while the discriminative patterns are made explicit.

C. Score Normalization Using Logistic Regression

This section describes how to normalize the original matching scores sc^O using the discriminative patterns w^A and w^B extracted from sorted cohort scores. Each of the three sets $\{sc^O, w^A, w^B\}$ contains complementary discriminative information which can be aggregated to enhance the recognition accuracy. The information carried by these three sets can be fused by training a linear SVM [45] or by applying a logistic regression [46] to compute more discriminative weights on

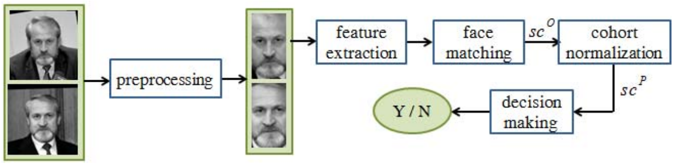

Fig. 6. Overview of the proposed face pair matching process applying the cohort score normalization.

each parameter. As shown in [37], logistic regression exhibits superior fusion performance over SVM. Therefore, a logistic regression is applied to approximate the final normalized scores according to the following expression:

$$sc^P = P\big(M \mid sc^O, w^A, w^B\big) \tag{5}$$

where $P\left(M \mid sc^O, w^A, w^B\right)$ is the conditional probability of (I^A, I^B) being a matching pair. A larger value of sc^P, implies a higher probability of I^A and I^B to be captured from the same subject.

IV. Application to Unconstrained Face Pair Matching

In this section, the application of the described picture-specific cohort score normalization process to unconstrained face recognition is discussed in detail. The LFW database is used to benchmark the algorithm performance.

A. The LFW Database

The Labeled Faces in the Wild dataset was composed by collecting from the web more than 13,000 face images, mostly from celebrities and news media. Two evaluation protocols are provided along with the dataset: image-restricted and unrestricted. In this paper, the performance of the subject-specific cohort normalization is evaluated on View 2, under the restricted setting. The 6,000 image pairs included in this subset, are divided into 10 splits, where the proportion of matching and non-matching pairs is balanced (1:1 ratio). As such, each split contains 300 matching and 300 non-matching pairs. The algorithm performance is measured by a 10-fold cross validation procedure [12]. Three versions of the LFW dataset are available: original, funnelled and aligned. The aligned version LFW-a, where faces are aligned with an unpublished algorithm, is employed in all experiments presented.

B. Face Pair Matching

The general process followed in the proposed normalization and matching approach is presented in Fig. 6. Four main steps are involved: *preprocessing, feature extraction, cohort normalization* and *decision making*.

1) Preprocessing: In the LFW aligned version, all the images are of the same size 250 × 250 pixels. At the pre-processing step, each image is simply cropped to remove the background, leaving only a face area of 150×80 pixels. At this stage no photometric normalization is performed.

2) Feature Extraction: Four facial descriptors are extracted from the face images: raw intensity, Gabor [6], LBP [7] and SIFT [8].

The first feature vector of length 12,000 is formed by concatenating all the raw intensity values of the pixels.

To compose the LBP feature vector, each image is first divided into non-overlapping blocks of size 10×10 pixels, and a 59-bin uniform LBP histogram is computed for each block. By concatenating the histograms computed for all blocks, a feature vector of length 7,080 is obtained.

To compose the Gabor feature vector, each image is filtered with a Gabor filter bank at five scales and eight orientations. The final Gabor feature vector is obtained by concatenating the responses at different pixels, uniformly selected with a 10×10 down-sampling rate. The resulting Gabor feature vector is of length 4,800.

To compose the SIFT feature vector, each image is divided into non-overlapping blocks of size 16 × 16 pixels. A 128 dimension SIFT descriptor is computed for each block. All SIFT descriptors are concatenated into a single vector of length 5,760.

3) Cohort Normalization: The matching scores are obtained by computing both the Euclidean distance and the Hellinger distance between two feature vectors. As described in [37], the degree of the polynomial used to compute the polynomial regression has little impact on the generalization performance. For simplicity, a linear function is employed to fit the two cohort score functions $f^A(h)$ and $f^B(h)$. To perform the logistic regression, an l_2-penalized logistic regression is computed which corresponds to the maximum likelihood estimate.

4) Decision Making: After cohort normalization, a threshold can be applied to the normalized scoresto achieve a final decision. As the normalized score is the probability of the two given samples to be a matching pair, generally a threshold with a value equal to 0.5 is set. Whenever the recognition accuracy is reported as a measure of the algorithm performance, a threshold of 0.5 is used. Whenever the Equal Error Rate is used as a performance measure, the corresponding threshold is the unique operating point where the False Accept Rate (FAR) is equal to the False Reject Rate (FRR) [17].

C. Experimental Results

In order to determine the actual performance of the proposed unconstrained face recognition approach, several experiments are presented, all performed on the LFW-a database.

1) Results From Individual Facial Descriptors: The first set of experiments is designed to test the improvement in classification accuracy by applying the cohort score normalization and adopting individual facial descriptors. For each of the 10 folds of LFW View 2, one out of the 10 splits is reserved as the cohort split, one split as the validation set, and the remaining eight splits for training the logistic regression weights. It is worth noting that different cohort splits are used in all 10 experiments.

For any of the 10-fold experiments, each dataset split is composed of 600 image pairs, for a total of 1,200 face images. In order to speed up the computation, only 600 randomly

TABLE I
COMPARATIVE CLASSIFICATION ACCURACY OF DIFFERENT DESCRIPTORS AND DISTANCES WITH AND WITHOUT COHORT NORMALIZATION

	Intensity	Gabor	LBP	SIFT
Euclidean	0.6502	0.6985	0.6500	0.7140
Euclidean with cohort	**0.6830**	**0.7560**	**0.7443**	**0.7703**
Hellinger	0.6497	0.7100	0.7132	0.7183
Hellinger with cohort	**0.6913**	**0.7680**	**0.7707**	**0.7738**

selected images from the cohort split (out of the 1,200 available) are used to compose the final cohort set.[2]

The obtained experimental results are described in Table I. The recognition accuracy for each feature type is reported as computed from the Euclidean distance and the Hellinger distance of the feature vectors. As shown in Table I, the cohort normalization allows an improvement of about 6% over the single Euclidean distance. By employing the LBP descriptor, an increase in accuracy of almost 9.5% is obtained. By using the Hellinger distance the accuracy is improved of about 5%. The highest accuracy (77.38% correct match) is achieved by applying the cohort normalization and the Hellinger distance on SIFT feature vectors.

Table I shows the absolute improvement in recognition accuracy due to the cohort normalization. However, to better evaluate the impact of the cohort normalization, the relative improvement of a given matching setting by using cohort normalization, is computed. Since there are 8 independent experiments (4 facial descriptors and 2 distances), the results are summarised as the relative change of the EER with respect to the performance of the baseline system [37], [47]. The EER is chosen as measure performance due to its sensitivity to minute changes induced by cohort score normalization. The relative change of the EER is computed as:

$$\text{rel. change of EER} = \frac{\text{EER}_{cohort} - \text{EER}_{baseline}}{\text{EER}_{baseline}} \tag{6}$$

where EER_{cohort} is the EER of a given system where cohort normalization is applied, whereas $\text{EER}_{baseline}$ is the EER scored by the same system but without cohort normalization. A negative change in the EER implies an improvement over the baseline system. Confidence intervals of the relative merit for each method with respect to the baseline system are also computed. These confidence intervals are plotted using a boxplot diagram, where the median, the first and third quarter as well as the fifth and 95-th percentiles of the data are plotted. The relative change of the EER for the described 8 individual experiments is shown in Fig. 7. As it can be noticed, in all 8 experiments, the cohort normalization process consistently improves the performance of the corresponding baseline system.

2) Comparison With the State-of-the-Art: In this section a comparison of the proposed cohort normalization-based approach with some of the state-of-the-art face recognition techniques is presented, through a series of experiments performed on the LFW database. Due to the large diversity among the information exploited by different algorithms (fusion of different descriptors, as well as the application of different distance metrics), it may be difficult to compare the performances with some algorithms in the literature. Therefore, the One-Shot Similarity (OSS, the best performing algorithm reported in [19]) and the Logistic Discriminant Metric Learning (LDML) [24] algorithms, using the same descriptors and also the same distance metrics, are chosen to be compared against the proposed cohort normalization-based system.

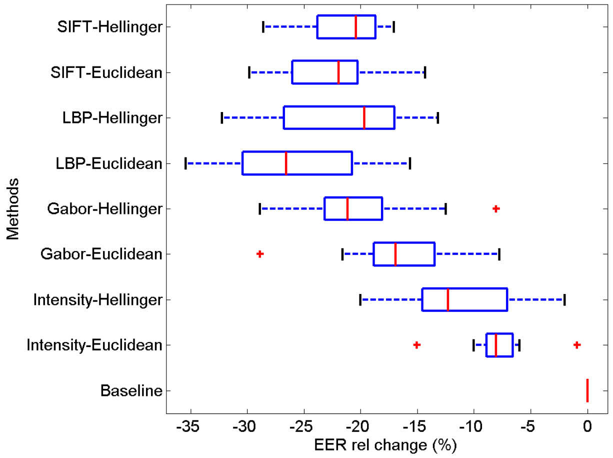

Fig. 7. Boxplot of the relative change of the EER using different descriptors and distance measurements.

TABLE II
COMPARATIVE CLASSIFICATION ACCURACY ON THE IMAGE-RESTRICTED BENCHMARK ("VIEW 2")

Algorithms	Euclidean	Hellinger
Gabor(C1)-OSS	0.7396	0.7437
LBP-OSS	**0.7663**	**0.7820**
SIFT-OSS	0.7576	0.7597
SIFT-LDML-PCA(35)	0.7660	**0.7750**
SIFT-LDML-PCA(55)	0.7280	0.7280
Gabor-Cohort	**0.7560**	**0.7680**
LBP-Cohort	0.7443	0.7707
SIFT-Cohort	**0.7703**	**0.7738**

Table II reports the comparative results obtained from the image-restricted benchmark ("View 2"). The image descriptors adopted are those reported in the original literature for the considered algorithms [19], [24]. The cohort score normalization-based system with the Gabor and SIFT features outperforms OSS when using either the Euclidean or the Hellinger distance measure. When adopting the LBP feature, the cohort-based approach performs slightly worse than the OSS algorithm. The accuracy reached by LDML on SIFT features, with PCA of dimension 35, is comparable with the cohort normalization-based algorithm. However, when the dimension of PCA is increased up to 55, the performance of LDML considerably decrease.

V. UNDERSTANDING THE COHORT BEHAVIOR

Even though much research efforts have been devoted to exploit useful information from a cohort/background dataset

[2] The term "cohort split" is used to represent the dataset split from which the cohort samples are selected, while "cohort set" represents the final fixed cohort set applied for score normalization. This corresponds to the set C described in section III-A. C^A and C^B are the "cohort lists", obtained by ordering the cohort samples.

for unconstrained face recognition, a limited knowledge has been acquired on the cohort behavior. For example, in most research papers, the authors randomly select a set of face images from one or more face databases to compose the cohort set. To the best of our knowledge, no attempt was ever made to understand and describe how to compose an optimal background set for a given face recognition task. To achieve a proper understanding of the cohort behavior, a set of experiments on both face pair matching and face verification are designed and performed. For face pair matching, the proposed picture-specific cohort normalization-based algorithm is applied, while for face verification, the polynomial regression-based cohort normalization proposed in [37] is adopted. As both cohort normalization algorithms extract discriminative information from cohort samples, the analysis of the experimental results will allow to determine the discriminative cohort behavior.

A. What is the Contribution of the Cohort Set in Matching Faces?

In this section the assumption that sorted cohort scores of face matching and non-matching pairs imply discriminative patterns, is verified by performing several experiments on the LFW database. The experimental setting is similar to that described in section IV-B. However, due to the limited number of face pairs, in this case the cohort scores are computed only for the eight development splits. For each face pair, two picture-specific cohort score profiles sc^A and sc^B, vectors of length 600, are obtained. It is worth noting that the ordering of the cohort score profile for I^A is determined by I^B; and that of I^B is determined by I^A. Finally, a total of 48,000 (= 2 × 300 × 8 × 10) matching cohort score profiles and 48,000 non-matching cohort score profiles are computed. The same experimental setting is applied for both the following qualitative and quantitative analysis.

1) Qualitative Analysis of the Cohort Discriminative Information: In order to qualitatively determine the discriminative information in the cohort dataset, the mean and standard deviation of matching and non-matching cohort score profiles are computed. The distributions obtained using the Gabor and LBP features with Euclidean distance are shown in Fig. 8 and Fig. 9. It can be readily observed that the two distributions have a different behavior. In fact, the cohort score profiles of matched face pairs show a steepest slope than the profiles of non-matched face pairs. This implies that the scores of matched face pairs increase with the rank order. This qualitative observation verifies the assumption that the cohort score profiles, sorted by the reciprocal face image, contain some discriminative information.

2) Quantitative Analysis of the Cohort Discriminative Information: The quantitative evaluation of the discriminative power for face pair matching, embedded in the sorted sorted cohort scores alone, is based on the computation of the Equal Error Rates. A comparison of the EERs, computed from the original matching scores and from the cohort discriminative patterns, is reported in Table III. The rows marked as "sc^O" report the EERs computed by using only the raw matching score, while the rows marked as "$w^A + w^B$" report the EERs computed by using only the discriminative patterns extracted from the sorted cohort scores. The latter EERs are obtained by means of a logistic regression, performed using only the two approximated parameters and without the raw score. The EERs computed from the cohort patterns is generally higher than those obtained by using the raw matching score. When using the raw intensity feature, both the matching and the cohort patterns report a very similar EER. In general, the values of the EERs are very similar, demonstrating that the cohort ordering

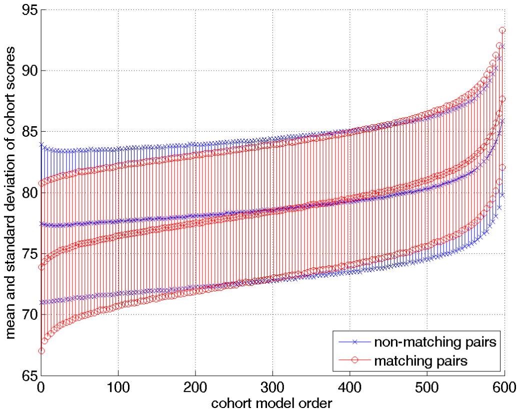

Fig. 8. Distribution of the cohort scores generated by the ordered cohort samples for matching and non-matching pairs using Gabor features.

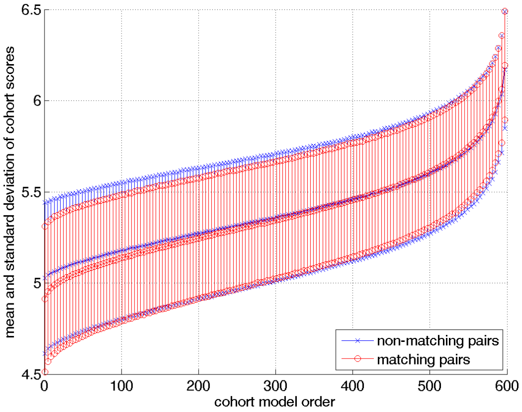

Fig. 9. Distribution of the cohort scores generated by the ordered cohort samples for matching and non-matching pairs using LBP features.

TABLE III
COMPARATIVE EERS COMPUTED FROM THE RAW SCORES AND THE COHORT DISCRIMINATIVE PATTERNS ALONE

	Intensity	Gabor	LBP	SIFT
sc^O (Euclidean)	**0.3453**	0.3047	0.3477	0.2980
$w^A + w^B$ (Euclidean)	**0.3603**	0.3717	0.3793	0.3557
sc^O (Hellinger)	**0.3480**	0.3000	0.2963	0.2927
$w^A + w^B$ (Hellinger)	**0.3417**	0.3667	0.3587	0.3580

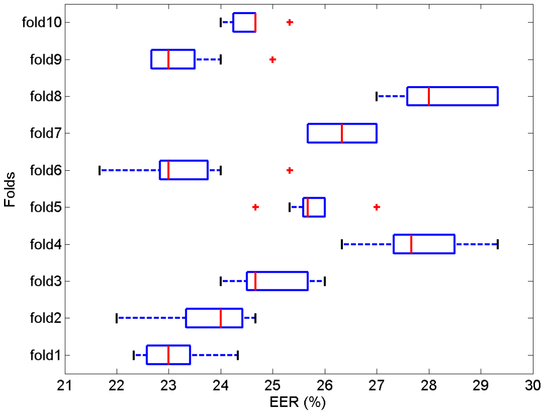

Fig. 10. Boxplot of the computed EERs with different choices of the cohort set using the Gabor features.

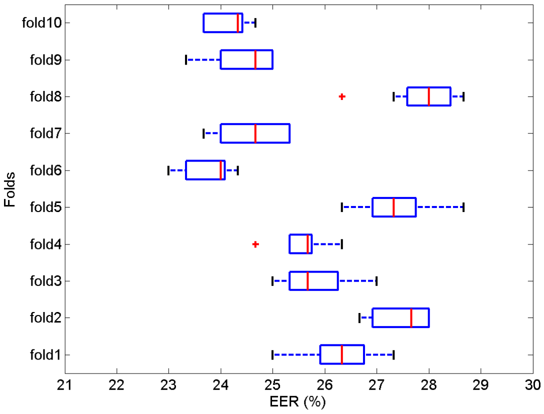

Fig. 11. Boxplot of the computed EERs with different choice of the cohort set using the LBP features.

itself conveys roughly the same *amount* of discriminative information of the raw matching scores.

B. How do the Choice and Size of the Cohort Dataset Affect Performance?

In all the experiments described above, the cohort set was randomly selected from a split. However, how the choice and size of any cohort split may impact the proposed cohort-based normalization procedure has not been considered. A further set of experiments is presented to evaluate the impact of these two parameters in the system performance.

1) Impact of the Choice of the Cohort Dataset: In the first set of experiments *different splits* from the data samples are used to compose the cohort set, but the size of the cohort dataset is left unchanged. It is worth noting that, with the exception of the evaluation split, for each fold experiment, one split out of a total of 9 can be used to compose the cohort set. This implies that each fold experiment can be performed 9 times, each time using a different cohort split. Given a cohort split, composed of 1,200 images, 600 images are subsampled from the set and included in the cohort set. A boxplot of the EERs, computed on the evaluation split, is used to illustrate the performance impact due to the change in the cohort sets. The results obtained by applying the Gabor and LBP descriptors and using the Euclidean distance are shown in Fig. 10 and Fig. 11. As it can be noticed, the choice of the cohort set involves a limited variation of the EER on the system performance.

2) Impact of the Cohort Dataset Size: In the second set of experiments the *size* of the cohort set is changed. In this case only one fold validation is performed on the same cohort split, but varying the number of face samples included in the cohort set from 100 to 900 images. Given a cohort split of $M = 1200$ images, m images are selected for the cohort set. For each value of m, 100 random samplings are performed, then the mean and standard deviation of the total 100 EERs are computed. The results obtained by applying the Gabor and LBP descriptors and using the Euclidean distance, are shown in Fig. 12. The solid lines represent the mean EERs, while the dashed lines represent the standard deviation of the EERs.

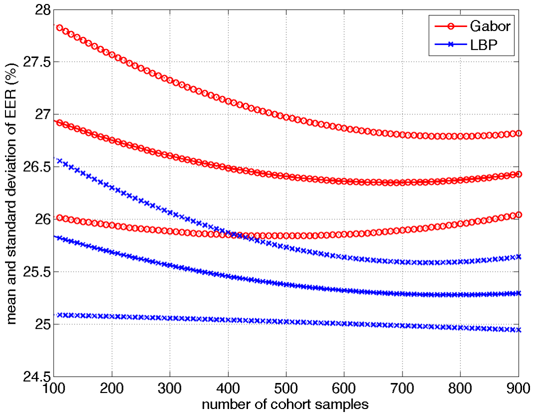

Fig. 12. Mean and standard deviation of EERs as the number of cohort samples increased.

As it can be noticed, the larger the size of the cohort set, the lower the EER. Also the standard deviation of the EER decreases with the mean. Therefore, increasing the number of cohort samples, the EER decreases up to a point, around 800 images, when the performance saturates and the EER slightly increases.

C. What is the Impact of the Quality of the Cohort Samples?

In the former experiment, the cohort set was composed from different splits of the same LFW dataset. As such, all the subsets of face samples had the same quality, thus producing similar cohort normalization performance. To better understand the impact of cohort's quality on the generalization performance, in this section, a set of experiments are performed on datasets composed of images of different quality. Both the FRGC ver2.0 database (with face verification protocols) and the LFW database (with face pair matching protocols) are used. With the FRGC ver2.0 database, the experiments are performed aiming to explore the impact of cohort quality on matching faces obtained in controlled environments (these will be referred as "lab faces"). On the other hand, by performing experiments with samples from the LFW database, the impact of cohort quality on matching faces

TABLE IV
EIGHT COMBINATIONS OF QUERY AND COHORT SAMPLES, VARYING THE QUALITY OF THE FACE IMAGES

Cohort condition	Good query	Bad query
Without cohort	Qgood	Qbad
Good cohort	QgoodCgood	QbadCgood
Bad cohort	QgoodCbad	QbadCbad
Joint cohort	QgoodCjoint	QbadCjoint

TABLE V
NUMBER OF CONTROLLED AND UNCONTROLLED IMAGES IN THE 5 FOLDS ON THE FRGC VER2.0 DATABASE

Fold No	1	2	3	4	5
controlled images	2,780	3,424	3,264	2,928	3,592
uncontrolled images	1,390	1,712	1,632	1,464	1,796

collected from real-world images (these will be referred as “wild faces”) will be analysed.

1) Impact on Matching Lab Faces: In the following experiments, all template models are acquired in well controlled conditions and are made from good quality face samples. On the other hand, the query samples are composed of both good and bad quality face samples. Three different cohort sets are assembled, composed of good quality, bad quality and joint face samples. The *good* cohort set is composed of face images captured in well controlled conditions. The *bad* cohort set is composed of bad quality, or uncontrolled, face samples. The *joint* cohort set is composed of both good and bad quality face images. In order to make a fair comparison, the three cohort sets are composed of the same number of images.

In total 8 possible combinations of good and bad query samples and good, bad and joint cohort sets are used to perform the experiments, as illustrated in Table IV. By denoting as “Q” the query set and “C” the cohort set, “Qgood” implies the direct comparison between the target and the good quality query samples, without cohort score normalization. “QgoodCgood” implies the matching of the target with the good quality query samples, with the cohort score normalization, using the cohort set made of good samples.

a) The FRGC ver2.0 database: The FRGC ver2.0 database [15] includes a testing protocol dividing the dat set into 6 different subsets or experimental data. In the following experiments, the face images from experiment 4 were used. The target set consists of 16,028 controlled still images, and the query set consists of 8,014 uncontrolled still images, captured from 466 subjects. 465 subjects are selected to perform a 5-fold cross validation experiment, by dividing the 465 subjects into 5 folds, each containing $465 \div 5 = 93$ different subjects. Finally, a total of 15,988 controlled images and 7,994 uncontrolled images are obtained. For each fold, the number of controlled images are listed together with that of uncontrolled images in Table V. For each of the 5-fold experiments, one fold is selected for the final evaluation, one fold for selecting cohort samples, and other three folds for development. In this way, the identities in the evaluation, development and cohort sets are disjoint from one another. Furthermore, for all 5 experiments, also the cohort folds are different from one another.

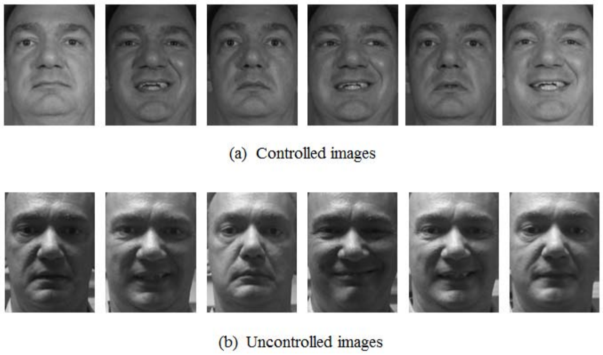

Fig. 13. Examples of aligned images from the FRGC ver2.0 database. (a) Good quality images. (b) Bad quality images.

b) Data configuration: As shown in Table V, for each fold, the number of controlled images is twice that of uncontrolled images. As shown in Table VI, the target and good query sets are composed from the controlled images, while the uncontrolled images are used to compose the bad query set. According to the FRGC testing protocol, each query sample is compared with all the target models in the target set. The total number of resulting matching tests, for each fold, is shown in Table VI. It is worth nothing that, for each fold, the same number of genuine and impostor matches between “Qgood” and “Qbad”, is obtained.

In order to compose the cohort set, the whole fold is first divided into three partitions: target, good query and bad query sets. Afterwards, 700 images are randomly selected from the good query set to build the good cohort set. The same procedure is followed to produce the bad cohort set from the uncontrolled images. Finally, half of the images from the good cohort set and half of the images from the bad cohort set are taken to build the joint cohort set. As a result, all three cohort sets are composed of the same number of images.

c) Normalization and feature extraction: All face images are geometrically normalized to a fixed size. From the provided coordinates of the four eye corners, the coordinates of the two eye centers are computed, and the distance between the eye centers is set to 60 pixels. Finally, all images are cropped to 110×80 pixels. Some sample normalised images are shown in Fig. 13. According to the setting described in section IV-B, Gabor feature vectors of length 3,520 and LBP vectors of length 5,192 are computed. The matching score is computed from the cosine similarity between two descriptors. A polynomial regression-based cohort normalization (l_2-penalized logistic regression, with a polynomial degree equal to 1), is applied to extract discriminative information from cohort samples [37].

d) Experimental results: The mean EERs of the 5 experiments are reported in Table VII. “Czero” represents the baseline system without cohort score normalization, i.e., the “Qgood” and “Qbad” matching listed in Table IV. The best performance is reported when matching good quality query samples and performing the normalization with good quality cohort samples. When matching bad query images, both the “Cgood” and “Cjoint” cohort normalization achieves similar

TABLE VI

DATA CONFIGURATION OF THE 5 FOLDS FOR THE VERIFICATION EXPERIMENT ON THE FRGC VER2.0 DATABASE

Fold	1	2	3	4	5
target models	1,390	1,712	1,632	1,464	1,796
good queries	1,390	1,712	1,632	1,464	1,796
bad queries	1,390	1,712	1,632	1,464	1,796
total matches	1,932,100	2,930,944	2,663,424	2,143,296	3,225,616
genuine matches	32,092	44,608	41,048	36,464	49,064
impostor matches	1,900,008	2,886,336	2,622,376	2,106,832	3,176,552

TABLE VII

MEAN EERS OF THE 5-FOLD EXPERIMENTS ON LAB FACE VERIFICATION WITH THREE DIFFERENT QUALITY COHORT SETS

Feature	Query	Czero	Cgood	Cbad	Cjoint
Gabor	Qgood	0.1123	**0.0586**	0.0853	0.0700
	Qbad	0.2867	0.2245	0.2658	**0.2122**
LBP	Qgood	0.0746	**0.0461**	0.0568	0.0497
	Qbad	0.3185	0.2330	0.2850	**0.2280**

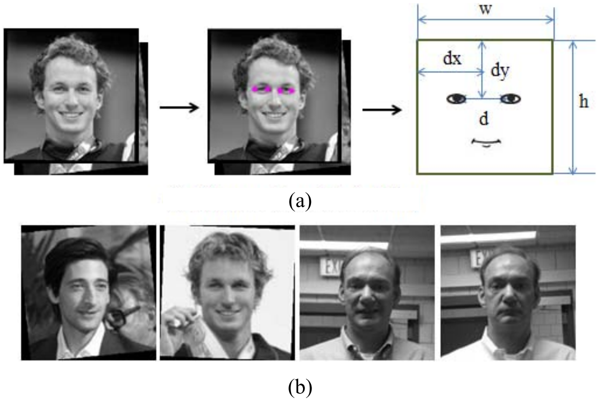

Fig. 14. Procedure for the lab face alignment. (a) Construction of the alignment template. (b) Sample wild faces and aligned lab faces.

TABLE VIII

MEAN EERS OF THE 10-FOLD EXPERIMENTS ON WILD FACE PAIR MATCHING WITH FOUR DIFFERENT COHORT SETS

Feature	Czero	Cgood	Cbad	Cjoint	Cwild
Gabor	0.3047	0.3020	0.2953	0.2857	**0.2537**
LBP	0.3477	0.2917	0.2857	0.2657	**0.2570**

performances. Cohort normalization provides worse performance when bad quality cohort samples are employed than using good quality cohort samples. For example, with either Gabor and LBP features, the “Cbad” normalization produces 4.13% and 5.20% higher EERs than using the “Cgood” normalization. As shown in Table VII, “QgoodCgood” achieves 5.37% and 2.85% lower EERs than the baseline system “Qgood” when matching Gabor and LBP features.

2) Impact on Matching Wild Faces: To study the impact of cohort quality on matching wild faces, a series of experiments on the LFW database are performed. The same experimental settings described in section IV-B are applied. For each of the 10-fold cross validation experiments, the cohort set is composed of 600 images taken from the split.

a) Lab cohort selection: In order to perform a balanced evaluation, along with the cohort sets made of images from the LFW dataset (“Cwild”), also “lab faces” from the FRGC ver2.0 database are employed. Similarly to the previous experimental setup, cohort sets of three different quality are employed: “Cgood”, “Cbad” and “Cjoint”, on the same 5 folds. For each of the 5 folds, 1,200 images are selected from the good query set and other 1,200 images from the bad query set. By equally partitioning the 1,200 good quality and the 1,200 bad quality datasets, two “Cbad”, “Cgood” and “Cjoint”, each composed of 600 images, are built. Each of the two “Cjoint” sets is composed of 600 images, where 300 images are randomly extracted from the “Cbad” set and other 300 images from the “Cgood” set. For each of the 5 folds 2 {“Cgood”, “Cbad”, “Cjoint”} sets are obtained, for a total of 10 cohort sets, which are applied to perform the cohort normalization for the 10-fold cross validation experiments on the LFW database. It is worth nothing that all cohort sets {“Cgood”, “Cbad”, “Cjoint”, “Cwild”}, for each of the 10-fold experiments, are made of different face samples.

b) Lab face alignment: In order to employ lab faces to help matching wild faces, lab faces are geometrically normalized to a common coordinate system derived from the wild faces as explained in Fig. 14(a). A set of nearly frontal face images are chosen from the LFW database, and a publicly available tool [48] is applied to automatically locate the four eye corners. The average distance between the eye centres (46 pixels) and the average coordinates of the midpoint between the eyes ([125, 113]) are used to geometrically normalize the lab faces. Two sample wild faces and aligned lab faces are shown in Fig. 14(b). The image cropping of the face area is performed according to the same procedure described in section IV-B1.

c) Experimental results: The mean EERs computed for the 10-fold experiments are reported in Table VIII. Differing from the result obtained from matching lab faces, Cohort normalization provides better performance when bad quality cohort samples are employed than using good quality cohort samples. When the “Cbad” and “Cjoint” sets are employed, the cohort normalization achieves slightly better performance than with “Cgood” cohort samples. As expected, the best performance is obtained by using wild samples to build the cohort set. For example, matching Gabor and LBP features, the “Cwild” normalization produces 4.83% and 3.47% lower EERs than using “Cgood”. These results suggest that, when matching wild faces, cohort samples selected from real-world images allows to achieve better performance than using cohort samples obtained under controlled environments.

TABLE IX
SIX COMBINATIONS OBTAINED FOR THE TARGET, QUERY AND COHORT SAMPLES

Cohort condition	Good query	Bad query
Without cohort	Qgood	Qbad
Good cohort 1	QgoodCms	QbadCms
Good cohort 2	QgoodCfs	QbadCfs

TABLE X
COHORT CONFIGURATION FOR THE 5 FOLDS EXTRACTED FROM THE FRGC VER2.0 DATABASE

	Fold No	1	2	3	4	5
Cms	cohort samples	694	716	702	718	712
	subjects	71	63	64	71	61
	min samples/subject	2	2	2	2	2
	max samples/subject	20	24	24	24	24
Cfs	cohort samples	694	716	702	718	712
	subjects	22	20	21	21	19
	min samples/subject	22	32	28	26	32
	max samples/subject	44	44	42	40	42

TABLE XI
MEAN EERS OF THE 5-FOLD EXPERIMENTS ON LAB FACE VERIFICATION WITH TWO DIFFERENT COHORT SETS

Feature	Query	Czero	Cms	Cfs
Gabor	Qgood	0.1123	0.0607	0.0618
	Qbad	0.2867	0.2240	0.2273
LBP	Qgood	0.0746	0.0471	0.0480
	Qbad	0.3185	0.2348	0.2337

D. How Many Subjects Should be Included in the Cohort Set?

The answer to the last question proposed in the introduction, a set of experiments is performed on the FRGC ver2.0 database. Given the results obtained from the previous experiments, only good quality cohort samples are employed for the cohort normalization, and two cohort sets are built. The first cohort set contains as many subjects as possible, each subject with the fewest possible samples. The second cohort set contains the fewest possible subjects, each subject with as many samples as possible. The term “Cms” refers to the first cohort set and “Cfs” to the second cohort set. The same 5 folds, as in the former experiments performed on the FRGC ver2.0 database, are used. The 6 resulting combinations of target, query and cohort samples are listed in Table IX. The overall testing configurations and experimental settings are the same described in section V-C.1, the only difference lies in the composition of the cohort sets. The configuration of the cohort sets for the 5 folds is shown in Table X. For each fold, “Cms” and “Cfs” are composed of the same number of face samples.

The results obtained in the experiments are summarised in Table XI. As it can be noticed, when matching either good or bad quality query samples, the cohort normalization performed using either “Cms” and “Cfs” achieves almost the same performance. Therefore, if the total number of sample faces is kept unchanged, no significant difference in performance is registered by changing the number of subjects in the cohort dataset.

VI. CONCLUSION

The recognition of faces under unconstrained conditions has been addressed. In order to facilitate the pairwise face matching a picture-specific cohort score normalization approach has been proposed. The adoption of a subject-specifc normalization process, which is naturally tailored on the unique features of the subject’s face, allowed to achieve better performance over traditional blind normalization processes.

This paper particularly provided a better understanding of discriminative cohort behavior addressing the following questions:

1) How much discriminative information is contained in the cohort samples alone?
2) How do the choice and the size of the cohort set affect the normalization performance?
3) What is the result of employing cohort samples of different quality?
4) Should a cohort set contain as many as possible subjects (each subject with the fewest possible samples) or as few as possible subjects (each subject with the utmost possible number of samples)?

It has been shown that the cohort information *alone* embeds a discrimination power which is just marginally worse than the raw matching score. When this information is properly extracted (with a polynomial regression), and appropriately combined with the raw matching scores (with logistic regression), an improvement in the face pairing is almost always achieved over the corresponding baseline system. This approach has been validated on the LFW database, achieving performance comparable with the current state of the art.

Concerning the quality and size of the cohort sample set, it has been shown that a larger cohort set size provides more stable and often better results up to a limit, when the performance saturates and even slightly degrades. On the other hand, cohort samples with different quality indeed produce different cohort normalization performance. Generally, when matching face images captured in a controlled environment, cohort samples of good quality (captured under controlled conditions) allow to achieve much better performance than bad quality (captured under uncontrolled conditions) cohort samples. However, whenever matching face samples captured in uncontrolled conditions, such as for the images in the Labeled Faces in the Wild dataset, good quality cohort samples produce much worse performance than bad quality samples. In contrast, using “wild”, uncontrolled cohort samples allows to achieve the best performance.

Regarding the number of subjects included in the cohort dataset, it has been shown that, if the image samples in the cohort set are all captured under controlled conditions, the number of subjects in the set has a marginal impact on the verification performance. The experimental results provided, as well as the conclusions drawn, may provide researchers with useful suggestions and hints for designing a suitable cohort/background set for score normalization in face recognition.

Whenever a biometric system operates under challenging conditions, cohort normalization can improve robustness and recognition accuracy. However, to achieve a thorough understanding of the background dataset behavior, more research efforts are required by using different background-based approaches. The regression-based algorithm could be bootstrapped to improve the computational speed. Even though the described subject-specific cohort normalization has been developed for single image holistic face matching, it can be further extended to still-to-video or video-to-video face recognition. Further research can be devoted to extended the proposed method to component-based face recognition to augment robustness to facial occlusions. The proposed method could be also extended to address other biometric modalities by exploiting modality-specific cohort datasets.

Acknowledgment

The authors would like to express sincere gratitude to Dr. Chi-Ho Chan, Department of Electrical Engineering, University of Surrey, for his valuable suggestions on the experiments. Sincere thanks would be also given to the reviewers for their detailed review, valuable comments and useful suggestions.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[2] N. Ramanathan and R. Chellappa, "Face verification across age progression," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3349–3361, Nov. 2006.

[3] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyezdan, "Plastic surgery: A new dimension to face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 441–448, Sep. 2010.

[4] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[5] Y. Xu, A. Zhong, J. Yang, and D. Zhang, "LPP solution schemes for use with face recognition," *Pattern Recognit.*, vol. 43, no. 12, pp. 4165–4176, Dec. 2010.

[6] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[7] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 469–481.

[8] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of SIFT features for face authentication," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2006, pp. 35–41.

[9] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui, "FRVT 2006: Quo vadis face quality," *Image Vis. Comput.*, vol. 28, no. 5, pp. 732–743, 2010.

[10] G. Hua *et al.*, "Introduction to the special section on real-world face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1921–1924, Oct. 2011.

[11] S. Z. Li, *Encyclopedia of Biometrics*. New York, NY, USA: Springer-Verlag, 2009.

[12] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces in 'Real-Life' Images, Detection, Alignment, Recognit.*, Marseille, France, Oct. 2008, pp. 1–14.

[13] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proc. Int. Conf. Spoken Lang. Process.*, Banff, AB, Canada, Oct. 1992, pp. 599–602.

[14] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1996, pp. 81–84.

[15] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 947–954.

[16] Y. Sun, M. Tistarelli, and N. Poh, "Picture-specific cohort score normalization for face pair matching," in *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2013, pp. 1–8.

[17] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Berlin, Germany: Springer-Verlag, 2008.

[18] (2014). *LFW Results*. [Online]. Available: http://vis-www.cs.umass.edu/lfw/results.html

[19] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.

[20] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2707–2714.

[21] Z. Lei and S. Z. Li, "Learning discriminant face descriptor for face recognition," in *Proc. 11th Asian Conf. Comput. Vis.*, 2012, pp. 748–759.

[22] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.

[23] G. Sharma, S. U. Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 1–12.

[24] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 498–505.

[25] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vis.*, 2011, pp. 709–720.

[26] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3554–3561.

[27] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2408–2415.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Describable visual attributes for face verification and image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1962–1977, Oct. 2011.

[29] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 497–504.

[30] S. Liao, A. K. Jain, and S. Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193–1205, May 2013.

[31] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3499–3506.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, Jan. 2000.

[33] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.

[34] N. Poh and J. Kittler, "On the use of log-likelihood ratio based model-specific score normalisation in biometric authentication," in *Advances in Biometrics*. Berlin, Germany: Springer-Verlag, 2007, pp. 614–624.

[35] W. D. Marslen-Wilson, "Functional parallelism in spoken word-recognition," *Cognition*, vol. 25, nos. 1–2, pp. 71–102, 1987.

[36] G. Aggarwal, N. K. Ratha, and R. M. Bolle, "Biometric verification: Looking beyond raw similarity scores," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, pp. 31–36.

[37] A. Merati, N. Poh, and J. Kittler, "User-specific cohort selection and score normalization for biometric systems," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 1, pp. 1270–1277, Aug. 2012.

[38] G. Aggarwal, N. K. Ratha, R. M. Bolle, and R. Chellappa, "Multi-biometric cohort analysis for biometric fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar./Apr. 2008, pp. 5224–5227.

[39] Y. Sun, C. B. Fookes, N. Poh, and M. Tistarelli, "Cohort normalization based sparse representation for undersampled face recognition," in *Proc. Asian Conf. Comput. Vis. Workshops*, Daejeon, Korea, Nov. 2012, pp. 1–13.

[40] A. Martínez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. 24, 1998.

[41] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 42–54, 2000.

[42] S. Tulyakov, Z. Zhang, and V. Govindaraju, “Comparison of combination methods utilizing T-normalization and second best score model,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–5.

[43] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore, “Activation of the middle fusiform ‘face area’ increases with expertise in recognizing novel objects,” *Nature Neurosci.*, vol. 2, no. 6, pp. 568–573, 1999.

[44] M. Bicego, G. Brelstaff, L. Brodo, E. Grosso, A. Lagorio, and M. Tistarelli, “Distinctiveness of faces: A computational approach,” *ACM Trans. Appl. Perception*, vol. 5, no. 2, pp. 1–18, May 2008.

[45] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[46] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes,” in *Advances in Neural Information Processing Systems*, vol. 14, T.G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA, USA: MIT Press, 2002, pp. 841–848.

[47] N. Poh and S. Bengio, “Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication,” *Pattern Recognit.*, vol. 39, no. 2, pp. 223–233, Feb. 2005.

[48] M. Everingham, J. Sivic, and A. Zisserman, “‘Hello! my name is... Buffy’—Automatic naming of characters in TV video,” in *Proc. 17th Brit. Mach. Vis. Conf.*, Edinburgh, U.K., Sep. 2006, pp. 1–10.

Massimo Tistarelli (SM’08) received the Ph.D. degree in computer science and robotics from the University of Genoa, Genoa, Italy, in 1991. He is currently a tenured Full Professor of Computer Science and the Director of the Computer Vision Laboratory with the University of Sassari, Sassari, Italy.

He has been a Project Coordinator and Task Manager in several projects on computer vision and biometrics funded by the European Community since 1986, and the Director of the Computer Vision Laboratory with the Department of Communication, Computer and Systems Science, University of Genoa, since 1994. He is currently with the University of Sassari, where he is leading several national and European projects on computer vision applications and image-based biometrics. He is a Founding Member of the Biosecure Foundation, which includes all major European Research Centers working in biometrics. He is the Chair of the Management Committee of the European Union COST Action IC1106 “Integrating Biometrics and Forensics for the Digital Age.” His main research interests cover biological and artificial vision (in particular, recognition, 3-D reconstruction, and dynamic scene analysis), pattern recognition, biometrics, visual sensors, robotic navigation, and visuomotor coordination. He has coauthored over 100 scientific papers in peer-reviewed books, conferences, and international journals. He is the Principal Editor of a book entitled *Handbook of Remote Biometrics* (Springer, 2009). He is one of the world-recognized leading researchers in biometrics. He is an Associate Editor of IEEE T-PAMI, *Pattern Recognition Letters*, and *Image and Vision Computing*. He is the Scientific Director of the Italian Platform for Biometric Technologies (established from the Italian Ministry of the University and Scientific Research), the first Vice President of the International Association for Pattern Recognition (IARP), the President of the Italian Chapter of the IEEE Biometrics Council, a member of the Conference Committee of the IEEE Biometrics Council, and a Fellow Member of IAPR.

Yunlian Sun received the M.E. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2010, and the Ph.D. degree in computer science from the University of Bologna, Bologna, Italy, in 2014. From 2010 to 2011, she was a Research Assistant with the Biometrics Research Center, The Hong Kong Polytechnic University, Hong Kong. After this appointment, she became a Ph.D. candidate in Ingegneria Elettronica, Informatica e delle Telecomunicazioni at University of Bologna, Italy from April 2011 to May 2014. The Ph.D. study was jointly conducted at both the University of Bologna and the University of Sassari, Sassari, Italy. During the study, she also spent several months with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the University of Warwick, Coventry, U.K., as a visiting student. Her research interests focus on biometrics, pattern recognition, and image processing, especially on face recognition.

Norman Poh (M’06) joined the Department of Computing as a Lecturer in Multimedia Security and Pattern Recognition in 2012. He received the Ph.D. degree in computer science from the Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland, in 2006. He was a Research Fellow with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Surrey, U.K., and a Research Assistant with the Idiap Research Institute, Martigny, Switzerland. His research interests focus on developing and applying pattern recognition theories to biometrics, information fusion, and healthcare informatics. He has authored over 80 peer-reviewed publications.

He received two personal Fellowships from the Swiss National Science Foundation (Young Prospective and Advanced Researcher Grants), and authored five Best Paper Awards, such as AVBPA 2005, ICB 2009, HSI 2010, ICPR 2010, and *Pattern Recognition Journal* in 2006. He was a recipient of the Researcher of the Year Award in 2011 from the University of Surrey. His project Exo-Brain received several prizes in the ICC 2013.

Dr. Poh is currently an Associate Editor of the *IET Biometrics Journal*, the IEEE Certified Biometrics Professional and Trainer, and a member the International Association for Pattern Recognition and the Education Committee of the IEEE Biometric Council.