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—Face recognition is one of the most widely used biometric systems due to its non-intrusive, natural and easy to use characteristics. However, automatic face recognition becomes very challenging whenever the acquisition conditions are unconstrained. In order to enhance the robustness of face recognition under challenging conditions, this paper proposes to adopt a cohort-based score normalization procedure. Specifically, the polynomial regression-based cohort normalization is extended to the unconstrained face pair matching problem. Extensive experiments conducted on the LFW benchmark demonstrate the effectiveness of cohort normalization on this challenging scenario. Furthermore, we advance the state of the art in cohort normalization by providing a better understanding of its discriminative cohort behavior. In particular, we find that the cohort information alone has a certain discriminative power which is just marginally worse than the raw matching score. A larger cohort set size gives more stable and often better results to a point before the performance saturates and slightly reduces. Finally, the experimental results on both the FRGC ver.2.0 database (lab face verification) and the LFW database (wild face pair matching) show that cohort samples of different quality indeed produce different cohort normalization performance. Generally, for matching faces captured under lab environments, using cohort samples of good quality leads to much better performance than using bad cohort samples. However, for matching wild faces, using wild cohort achieves the best performance.

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

I. INTRODUCTION

PEOPLE 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., different illumination conditions), facial expression, the interaction with the face acquisition device (causing different 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 mainly been devoted to recognizing faces collected in controlled lab environments. A number of face databases have been assembled to understand the effect 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 [4] and manifold [5] learning methods are two representative holistic methods. 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 the unconstrained setting has two important applications: surveillance and semantic web search. With the popularity and increasing number of video surveillance cameras installed in public venues, face recognition is no doubt an important instrument in the fight against crime. At the same time, social web sites such as Picasa and Facebook have generated an unprecedented volume of photos and videos.

Automatic face recognition will play an increasingly important role in understanding these media contents by solving the identity recognition problem. For instance, photo tagging is a convenient feature that allows the end user to retrieve photos of friends and family quickly. The face images in the above applications 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.

In this paper, we will consider the face pair matching problem and will use the Labeled Faces in the Wild (LFW) database [12]. It 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 an enormous 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.

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 Department of Computing, University of Surrey, Guildford, Surrey, GU2 7XH, UK e-mail: n.poh@surrey.ac.uk
Manuscript received Month Day, Year; revised Month Day, Year.
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 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 are as follows:

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 questions are addressed:

A. How much discriminative information is contained in the cohort samples alone? This discriminative information is 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 if we use cohort samples of different quality for score normalization?

D. Should a cohort set contain subjects as many as possible (each subject with the fewest possible samples) or as few as possible (each subject with the utmost possible number of samples)?

II. RELATED WORK

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

A. Unconstrained Face Recognition

Since its release in 2008, the LFW database has received a considerable attention. Several algorithms were developed, specifically for handling large amounts of real-world face images [18]. Among these algorithms, there are quite a few focusing 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 (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]. Liu et al. [30] proposed an alignment-free sparse representation approach for partial face recognition. The gallery descriptors used in this approach 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] on the spatial-appearance features.

B. Cohort Score Normalization

In a typical biometric verification system, the decision on the identity of the biometric sample is made directly based on the matching score between the query sample and the claimed template model. Due to various forms of noise in the data capturing, 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-norm (T-norm) [41] are the two most common algorithms adopted in practical biometric applications. Both techniques assume that score distribution of each subject class follows Gaussian distribution, thus use the first and second order moments of cohort scores for distribution scaling. However, the cohort scores used in Z-norm are matching scores between the template and cohort samples, while those in T-norm are matching scores between the query and cohort samples. On the other hand, methods proposed in [37], [38], [42] attempt to exploit the patterns of sorted cohort scores. Among these approaches, 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. Note that this technique performs not only the matching between the query sample and the cohort samples during the test stage, but also the matching between each enrolled template and the cohort samples during the enrollment stage for producing the user-specific cohort rank orders.

III. PICTURE-SPECIFIC COHORT SCORE NORMALIZATION

In principle, a subject-specific face representation allows to maximize the discrimination capability for each individual. This approach requires to develop a computational model which embodies information which is peculiar for each subject. 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. The subject-based model is important because the hypothesis space, made of the score distributions subjecting to being 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 one.

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

The general idea is to apply a polynomial regression-based cohort normalization to face pair matching under uncontrolled environments so that it can well normalize against score distribution variations between the two pictures presented in a given pair of images. Fig. 4 illustrates the proposed approach.

A. Picture-Specific Cohort Selection

As presented in [37], in a verification system, the distribution of cohort scores obtained by matching a number of impostor and genuine query samples with cohort samples, which are first sorted with respect to their similarity to the claimed template, exhibits a discriminative pattern. Consequently, it is reasonable to assume that in the pair matching scenario, sorted cohort scores of matching pairs and non-matching pairs imply
similar discriminative patterns. We will verify this assumption in our latter experiments.

Let \((I^A, I^B)\) denote a pair of images to be compared and \(sc^D\) be the raw matching score between \(I^A\) and \(I^B\). Given an additional fixed cohort set \(C = \{c_1, ..., c_h, ..., c_H\}\), in which there are totally \(H\) cohort samples, we need to select an exclusive cohort list for each of \(I^A\) and \(I^B\). We denote the two picture-specific cohort lists as \(C^A\) and \(C^B\) respectively. Both \(C^A\) and \(C^B\) are sorted variants of \(C\), the only difference among the three sets lies in the rank order of cohort samples.

Before the selection of \(C^A\) and \(C^B\), a set of cohort scores between each picture of the pair and all the cohort samples in \(C\) are first computed. Cohort samples sorted with respect to their closeness to \(I^B\) constitute \(C^A\). Let \(\{c_1^A, ..., c_h^A, ..., c_H^A\}\) be 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 we can determine the cohort list for picture \(I^B\), that is \(\{c_1^B, ..., c_h^B, ..., c_H^B\}\), in which \(c_1^B\) is the most similar cohort sample to \(I^A\). Next we can get two picture-specific cohort score lists \(sc^A = \{sc_1^A, ..., sc_h^A, ..., sc_H^A\}\) and \(sc^B = \{sc_1^B, ..., sc_h^B, ..., sc_H^B\}\). The \(H\) scores in \(sc^A\) are matching scores between \(I^A\) and each cohort sample in \(C^A\), hence \(sc_h^A\) is the matching score between \(I^A\) and \(c_h^A\).

Hence in our approach, we have two picture-specific cohort score lists, which are generated using cohort samples sorted by the reciprocal image. This is different from the approach proposed in [37], where there is only one user-specific cohort score list which is generated by cohort samples sorted with respect to their closeness to the claimed template. With the two cohort score lists \(sc^A\) and \(sc^B\), next we focus on how to extract the discriminative patterns involved in them using 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. Specifically, with the two picture-specific cohort score lists \(sc^A\) and \(sc^B\) obtained above, we can get the following two functions of rank orders.

\[
sc_h^A = f^A (h) \quad (1)
\]

\[
sc_h^B = f^B (h) \quad (2)
\]

By polynomial regression, the noise is greatly weakened. As shown in Fig. 5, the fitted curves (i.e., \(w^A\) and \(w^B\)) in Fig. 5. Note that the polynomial degree is set to 1, i.e., we use linear function to fit the fitted curves (i.e., \(w^A\) and \(w^B\)) in Fig. 5. The fitted curves clearly show the discriminative patterns included in sorted cohort scores.

To validate the effectiveness of \(w^A\) and \(w^B\) on distinguishing matching pairs from non-matching pairs, we show the cohort score profiles (i.e., \(sc^A\) and \(sc^B\)) of a matching pair and a non-matching pair from the LFW database as well as the fitted curves (i.e., \(w^A\) and \(w^B\)) in Fig. 5. Note that the polynomial degree is set to 1, i.e., we use linear function to fit cohort score profiles. As can be observed, the cohort profiles versus cohort rank order are very noisy. The discriminative information between the cohort profiles of matching pair and those of non-matching pair is not significant. By polynomial regression, the noise is greatly weakened. As shown in Fig. 5, the fitted curves clearly show the discriminative patterns.

### C. Score Normalization Using Logistic Regression

In this section, we show how to normalize the original matching score \(sc^O\) using the discriminative patterns \(w^A\) and \(w^B\) extracted from sorted cohort scores. Each of the three components \(\{sc^O, w^A, w^B\}\) contains different discriminative power and should be aggregated in a reasonable way. We can fuse these parameters by training a linear SVM [44] or logistic regression [45] to obtain more discriminative weights on each parameter using a development data set. As observed in [37], logistic regression shows the superiority of fusion performance over SVM on the face modality. Therefore, here
we use logistic regression to approximate the final normalized score as follows:

\[ sc^P = P(M \mid sc^O, w^A, w^B) \]  

(5)

where \( P(M \mid sc^O, w^A, w^B) \) represents the probability of being a matching pair. To put it differently, the larger \( sc^P \) is, the more probable \( I^A \) and \( I^B \) come from the same subject.

IV. APPLICATION TO UNCONSTRAINED FACE PAIR MATCHING

In this section, we show how picture-specific cohort score normalization can be applied to face pair matching on the LFW database in detail.

A. The LFW Database

In this data set, more than 13,000 face images were collected from the web. There are two evaluation settings provided by the authors of the LFW: image-restricted and unrestricted. In this paper, we evaluate performance on View 2 under the restricted setting, which contains 6,000 image pairs. The image pairs are divided into 10 splits in such a way that the proportion of matching and non-matching pairs is balanced (1:1 ratio). Therefore, each split contains 300 matching and 300 non-matching pairs. The performance of an algorithm is measured by a 10-fold cross validation procedure. For more details, refer to [12]. There are three versions of the LFW available: original, funneled and aligned. Here we use the aligned version in all experiments.

B. Face Pair Matching Framework

The framework of our approach is presented in Fig. 6. There are mainly four steps involved: preprocessing, feature extraction, cohort normalization and decision making.

1) Preprocessing: In the aligned version, all the images are of the same size 250 × 250 pixels. At the preprocessing step, we simply crop the image to remove the background, leaving only a face area of 150 × 80 pixels. Note that we do not apply any photometric approaches at this step.

2) Feature Extraction: For the feature extraction step, we carry out experiments using four facial descriptors: intensity, Gabor [6], LBP [7] and SIFT [8]. The intensity feature vector of length 12,000 is formed by concatenating all the pixels. For extracting LBP features, we first divide each image into non-overlapping blocks of 10 × 10 size and extract a 59-bin uniform LBP histogram for each block. After concatenating the histograms of all the blocks, we can get a feature vector of length 7,080.

To compute the Gabor feature, we adopt five scales and eight orientations of the Gabor filters. The final Gabor feature vector is obtained by concatenating the responses at different pixels selected uniformly with a 10 × 10 down-sampling rate. The length of the Gabor feature vector is 4,800. For the SIFT descriptor, we also divide the image into non-overlapping blocks of 16 × 16 size, and compute a 128D SIFT descriptor for each block. All descriptors are then concatenated into a single vector of length 5,760.

3) Cohort Normalization: The methodology of cohort score normalization has been detailed in the above sections. Here we present only some points which are related to the experimental design. For the matching score, we compute both the Euclidean distance and the Hellinger distance between two descriptors. As concluded in [37], the degree of polynomial has little impact on the generalization performance, for simplicity, we employ a linear function to fit the two cohort score functions \( f^A(h) \) and \( f^B(h) \), i.e., \( n = 1 \). For logistic regression, we use \( l_2 \)-penalized logistic regression which leads to maximum likelihood estimate.

4) Decision Making: After cohort normalization, we can directly use the normalized score for the final decision. To achieve this, a threshold needs to be first determined. As described above, the normalized score is the probability of being a matching pair. Hence generally, 0.5 is set as the threshold. Two images with a matching score higher than the threshold are classified into the matching pair, and vice versa for those with matching score lower than the threshold. In this paper, we use 0.5 as the threshold in experiments where classification accuracy is used as the performance evaluation measure. While for those experiments using EER, the threshold is the unique operating point where the False Accept Rate (FAR) is the same as the False Reject Rate (FRR) [17].

C. Experimental Results

In this section, we show the experimental results on the LFW database.

1) Results from Different Individual Descriptor: Our first set of experiments are designed to test the classification accuracy increased by cohort score normalization with individual facial descriptor. For each of the 10 folds of View 2, we reserve one out of the 10 splits as the cohort split, another split as the validation set, and the remaining eight splits for training logistic regression weights. Note that in the 10 experiments, the cohort splits are different from one another.

Recall that each split has 600 image pairs, or a total of 1,200 face images. In order to speed up the computation, we use only 600 randomly selected images from the cohort split (out of the 1,200 available) to constitute the final cohort set for any of 10-fold experiments. It is worth noting that, we use the term “cohort split” to represent the split where we choose cohort samples, while “cohort set” represents the final fixed cohort set for score normalization, i.e., the before mentioned C. For \( C^A \) and \( C^B \), we call them “cohort lists”, in which cohort samples are ordered.

The results are described in Table I. Each type of feature is tested with the original feature vector (Euclidean distance)
and the square root of the feature vector (Hellinger distance). As shown in Table I, cohort normalization improves about 6% over the Euclidean distance. With LBP descriptor, we can get an increased accuracy of almost 9.5%. Using the Hellinger distance improves the accuracy by about 5%. The highest accuracy we can get from a single type of feature is 0.7738 using cohort normalization with the Hellinger distance of SIFT descriptors.

Table I shows the absolute improvement introduced by cohort normalization. However, it would be more interesting if the relative improvement of a given system compared to its respective baseline that is one without using cohort normalization, is shown. Since there are 8 independent experiments (due to 4 facial descriptors and 2 distances), we shall summarize the results using the relative change of EER [37], [46] with respect to the performance of the baseline system. Here we employ EER as the performance measure due to its sensitivity to minute changes induced by cohort score normalization. The relative change of EER is given as follows:

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

where EER_{cohort} is the EER of a given system with cohort normalization whereas EER_{baseline} is that of the baseline (without cohort normalization). A negative change of EER implies an improvement over the baseline system. This statistic has the advantage that one can establish confidence intervals of the relative merit of our approach with respect to the baseline system. These confidence intervals can be conveniently visualized using a boxplot, where the median, the first and third quartile as well as the fifth and 95-th percentiles of the data can be visualized. The relative change of EER for the above 8 individual experiments is illustrated in Fig. 7. As can be observed, in all of the 8 experiments, our cohort normalization mechanism improves the baseline system significantly.

2) Comparison with the State-of-the-Art Methods: Next we compare our cohort normalization based approach with those state-of-the-art methods on the LFW database. Due to diverse information used by different algorithms (fusion of different descriptors as well as different metrics), it is more reasonable to make comparisons among methods using the same descriptors and also the same distance metrics. Here, we choose OSS (the best result of [19]) and LDML [24] as the control algorithms.

Table II reports the comparative results on the image-restricted benchmark (“View 2”). Note that here we show only several descriptors since only the results of these descriptors are provided in the original literature [19], [24]. In comparison with OSS, cohort score normalization with Gabor and SIFT features outperforms OSS when using either Euclidean or Hellinger as the distance measure. With LBP feature, our approach is slightly worse than OSS. When comparing with LDML using SIFT feature, the accuracy obtained by LDML with PCA of dimension 35 is comparable with cohort score normalization. However, when the dimension of PCA turns 55, the performance of LDML decreases notably, and is much worse than ours.

V. Understanding of Cohort Behavior

Though much work has been done in exploiting useful information from a cohort/background set for unconstrained face recognition, little is known about the understanding of cohort behavior. For example, in most literature, the authors randomly select a set of face images from one/more known face databases to form the cohort set. To our best knowledge, there is no literature working on how to organize a good background set for face recognition. Are there any suggestions we can follow when performing this task? To achieve a proper understanding of cohort behavior, in this section, we conduct a set of experiments on both face pair matching and face verification. For face pair matching, we still

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Euclidean</th>
<th>Hellinger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor(C1)-OSS</td>
<td>0.7396</td>
<td>0.7437</td>
</tr>
<tr>
<td>LBP-OSS</td>
<td>0.7663</td>
<td>0.7820</td>
</tr>
<tr>
<td>SIFT-OSS</td>
<td>0.7576</td>
<td>0.7597</td>
</tr>
<tr>
<td>SIFT-LDML-PCA(35)</td>
<td>0.7660</td>
<td>0.7750</td>
</tr>
<tr>
<td>SIFT-LDML-PCA(55)</td>
<td>0.7280</td>
<td>0.7280</td>
</tr>
<tr>
<td>Gabor-Cohort</td>
<td>0.7560</td>
<td>0.7680</td>
</tr>
<tr>
<td>LBP-Cohort</td>
<td>0.7443</td>
<td>0.7707</td>
</tr>
<tr>
<td>SIFT-Cohort</td>
<td>0.7703</td>
<td>0.7738</td>
</tr>
</tbody>
</table>
Fig. 8. Distribution of cohort scores generated by ordered cohort samples for matching and non-matching pairs using Gabor feature.

use our proposed picture-specific cohort normalization, while for face verification, we employ polynomial regression-based cohort normalization proposed in [37]. Note that both cohort normalization algorithms extract discriminative information from cohort samples, thus we are interested to explore the discriminative cohort behavior involved in cohort samples.

A. How Much Can Cohort Alone Help Matching Faces?

1) Discriminative Information Illustration: Now we verify our former assumption that in face pair matching scenario, sorted cohort scores of matching pairs and non-matching pairs imply discriminative patterns, using the LFW database. The experimental settings are similar to the first set of experiments on testing cohort score normalization with individual descriptor. However, here we compute cohort scores for only the eight development splits not for the single evaluation split due to its limited number of pairs. For each pair, we can get two picture-specific cohort score profiles $sc_A$ and $sc_B$, each of which is a single vector of 600. Recall 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, we can get a total of 48,000 ($= 2 \times 300 \times 8 \times 10$) matching cohort score profiles and 48,000 non-matching cohort score profiles. Next we respectively compute the mean and standard deviation of matching and non-matching cohort score profiles. Fig. 8 and Fig. 9 show the distributions using Gabor and LBP features.

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

The comparative EERs of using original matching scores and cohort discriminative patterns are listed in Table III. "$sc^O$" rows are EERs obtained by using only the raw matching score, while "$w^A + w^B$" rows list EERs which are obtained by using only the discriminative patterns extracted from sorted cohort scores. To realize this, logistic regression is performed using only these two approximated parameters without the raw score. With the Euclidean distance, using the discriminative patterns leads to EERs 4% higher than those obtained by using the raw matching score; while this value is 5% with the Hellinger distance. However, for the intensity feature, both the baseline and cohort scheme have very comparable EERs. Indeed, the cohort scheme is slightly better in performance, having attained 0.3417 compared to its baseline counter that has an EER of 0.3480.

B. How Do the Choice and the Size of the Cohort Set Affect the Performance?

In all of the above experiments, for each of the 10 experiments, we have selected the cohort set randomly from a split. However, we do not know how the choice of any cohort split, as well as the size of the cohort set, may have impact on the proposed cohort-based normalization procedure. For this reason, next we carry out a set of experiments to evaluate the impact of these two design parameters.

1) Impact of Different Choices: First, we use different splits as the cohort split but the size of the cohort set remains the...
same. Note that for each fold experiment, except the evaluation split, we can select one split from the other 9 splits for the cohort split. In other words, we can perform each fold experiment 9 times, each time using a different cohort split. With the cohort split (1,200 images), we still select only half images for the final cohort set (600 images). We use the boxplot of EERs to illustrate the impact caused by using different cohort sets for the 10 folds. The results of Gabor and LBP descriptors using the Euclidean distance are shown in Fig. 10 and Fig. 11. It can be observed that the choice of the cohort set introduces an acceptable variation of approximate 1% EER on the system performance (as measured on the evaluation split).

2) Impact of Different Sizes: Next, we change the size of the cohort set. Still on the LFW database, however, we do not consider all the 10 experiments this time but keep to only one fold due to the high computational cost, that is, we test only one split. Denote the number of images in the cohort split as $M$ (here $M = 1200$). We select only $m$ images for the cohort set. Let the value of $m$ vary from 100 to 900. This is a “$m$ choose $M$” problem and is a combinatorial problem. There are many choices of “$m$ choose $M$”. For each value of $m$, we run 100 random samplings, then compute the mean and standard deviation of the total 100 EERs. Fig. 12 shows the results of Gabor and LBP descriptors using the Euclidean distance. The solid lines represent the mean of EERs, while the dashed lines represent the standard deviation. As shown in this figure, the larger the size of the cohort set, the lower the EER. At the same time, the standard deviation of the EER will also decrease. In other words, with more cohort samples, one can get a more stable result to a point before the performance saturates and slightly reduces.

C. What Impact Will Be Brought about if Using Cohort Samples of Different Quality?

In the former experiment on exploring different choices of the cohort set, we determined the cohort set from different splits. As is known, the images in each of the 10 splits of the LFW database are all real-world images. Put it differently, the images in the 10 splits are of similar quality to one another. In such a way, cohort samples from different splits will lead to similar cohort normalization performance, which has been verified by the experiment presented above. To find out the effect of cohort’s quality on the generalization performance, in this section, we conduct a set of experiments on both the FRGC ver2.0 database (following face verification protocols) and the LFW database (following face pair matching protocols). In the FRGC ver2.0 database, the experiments are performed aiming to explore the impact of cohort quality on matching faces obtained under lab environments (We use “lab faces” to represent them). On the other hand, through experiments on the LFW database, we are interested to find out the impact of cohort quality on matching faces collected from real-world images (“wild faces”).

1) Impact on Matching Lab Faces: Generally, when building a template model for each user in the enrollment stage of a verification system, it is difficult to get a template model which captures all possible intra-personal variations. Therefore, in our following experiments, we assume all the template models
are acquired in well controlled conditions, thus having good quality. For the query samples, it is unknown whether they are obtained in controlled conditions or not. Hence, we consider two kinds of query samples: good quality and bad quality. For the cohort samples, we organize three sets of cohort samples: good, bad and joint cohort sets. In the good cohort set, all the cohort face images are captured in well controlled conditions. While in the bad cohort set, cohort samples are of bad quality. For the joint cohort set, we include cohort samples with both good and bad quality. As presented in the former experiment on exploring the size of the cohort set, cohort sets of different sizes produce different normalization performance. Hence, to make a fair comparison among the three cohort sets, we collect the same number of cohort samples in them.

If we use “T” to represent template models, “Q” to represent queries and “C” to denote cohort samples, we have a total of 8 combinations among template models, queries and cohort samples as shown in Table IV. “TQgood” is the direct comparison between the target and the query with good quality, i.e., without cohort score normalization. “TQgoodCgood” represents using cohort samples of good quality to normalize the raw matching score between the target and the good query.

The FRGC ver2.0 Database In the FRGC ver2.0 database [15], the authors provide the researchers with 6 experiments. For our experiments, we choose face images from the Experiment 4. This experiment is designed to measure progress on recognition from uncontrolled frontal still images. The target set consists of controlled still images, and the query set consists of uncontrolled still images. There are a total of 16,028 controlled images and 8,014 uncontrolled images included in the Experiment 4. These images are from 466 subjects. For our experiments, we choose 465 subjects to perform a 5-fold cross validation experiment by dividing the 465 subjects into 5 folds with each containing 465/5 = 93 different subjects. Finally, we have a total of 15,988 controlled images and 7,994 uncontrolled images. For each fold, we list the number of controlled images together with that of uncontrolled images in Table V. For each of the 5 folds, we choose one fold for the final evaluation, one fold for selecting cohort samples, and the other three folds for development. In this way, the identities in the evaluation, development and cohort sets are disjoint from one another. Furthermore, for the 5 experiments, we keep their cohort folds different from one another.

Table IV: Eight Combinations of Target, Query and Cohort Samples.

<table>
<thead>
<tr>
<th>Cohort condition</th>
<th>Good query</th>
<th>Bad query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without cohort</td>
<td>TQgood</td>
<td>TQbad</td>
</tr>
<tr>
<td>Good cohort</td>
<td>TQgoodCgood</td>
<td>TQbadCgood</td>
</tr>
<tr>
<td>Bad cohort</td>
<td>TQgoodCbad</td>
<td>TQbadCbad</td>
</tr>
<tr>
<td>Joint cohort</td>
<td>TQgoodCjoint</td>
<td>TQbadCjoint</td>
</tr>
</tbody>
</table>

Table V: Numbers of Controlled and Uncontrolled Images in the 5 Folds on the FRGC ver2.0 Database.

<table>
<thead>
<tr>
<th>Fold No</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># controlled images</td>
<td>2,780</td>
<td>3,424</td>
<td>3,264</td>
<td>2,928</td>
<td>3,592</td>
</tr>
<tr>
<td># uncontrolled images</td>
<td>1,390</td>
<td>1,712</td>
<td>1,632</td>
<td>1,464</td>
<td>1,796</td>
</tr>
</tbody>
</table>

Feature Extraction and Cohort Normalization Before performing feature extraction, all the images are geometrically normalized to a fixed size. With the provided coordinates of the four eye corners, we compute the coordinates of the two eye centers. After the location of eye centers, the distance between the eye centers is set as 60 pixels. Finally, all the images are normalized to the size of 110 × 80. Several normalized examples are shown in Fig. 13. For feature extraction, we work on Gabor [6] and LBP [7] features. Following the same settings to those presented in Section IV. B, we can get Gabor feature vectors of length 3,520 and LBP of length 5,192.
For the matching score, we compute the cosine similarity between two descriptors. For cohort normalization, we employ polynomial regression-based cohort normalization to extract discriminative information from cohort samples [37]. The polynomial degree remains 1. For logistic regression, we still use $l_2$-penalized logistic regression.

**Experimental Results** As described above, our experiment is a 5-fold cross validation experiment. We summarize the mean EERs of the 5 experiments in Table VII. We use “Czero” to represent the baseline system without cohort score normalization, i.e., the systems “TQgood” and “TQbad” in Table IV. Obviously, systems with cohort score normalization, no matter using what kind of cohort samples, outperform the baseline systems. Interestingly, we find that, for matching lab faces with cohort score normalization, the quality of cohort samples indeed matters. For recognizing query samples obtained under well controlled lab environments, using cohort samples of good quality achieves the best performance. For matching bad queries, using “Cgood” and “Cjoint” obtains the similar performance. With bad cohort samples, cohort normalization leads to much worse performance than with good cohort samples. For example, for Gabor and LBP features, using “Cbad” leads to 4.13% and 5.20% higher EERs than using “Cgood”. As shown in Table VII, with good cohort samples, “TQgoodCgood” gets 5.37% and 2.85% lower EERs than the baseline system “TQgood” for Gabor and LBP features. The figures for the comparison between “TQbadCgood” and “TQbad” are 6.22% and 8.55%.

2) **Impact on Matching Wild Faces**: To study the impact of cohort quality on matching wild faces, next we perform a series of experiments on the LFW database. We follow the same experimental settings to those in Section IV. B, however, employ only Gabor and LBP features. Recall that for each of the 10-fold cross validation experiments, we select from one split 600 images to construct the cohort set. For these cohort samples, we use “Cwild” to represent their quality due to the fact that they are selected from real-world images.

**Lab Cohort Selection** In addition, we employ also lab faces for the cohort use to help matching wild faces. The lab cohort samples used in the following experiments are chosen from the FRGC ver2.0 database. Similarly, we consider lab cohort of three different quality conditions: “Cgood”, “Cbad” and “Cjoint”. We use the same 5 folds as those used in the former experiments on the FRGC ver2.0 database. For each of the 5 folds, we select 1,200 images from the good query set and the same 1,200 images from the bad query set. Next, we choose half good query for one good cohort set (“Cgood” with 600 images) and the same half bad query for one bad cohort set (“Cbad” with 600 images). The corresponding joint cohort set consists of half “Cgood” images (300 images) and half “Cbad” images (300 images). In this way, all the good, bad and joint cohort sets contain 600 images. Similarly, the other half good query (1,200 − 600 = 600 images) and corresponding bad query (1,200 − 600 = 600 images) are used to construct another “Cgood” and “Cbad” sets. Half “Cgood” and half “Cbad” constitute their corresponding “Cjoint” set. In such a way, for each of the 5 folds, we can generate 2 {“Cgood”, “Cbad”, “Cjoint”}. That is, finally we have a total of 10 {“Cgood”, “Cbad”, “Cjoint”}, which can then be used to perform cohort normalization for the 10-fold cross validation experiments on the LFW database.

It is worth noting that, in the 10-fold experiments, the {“Cgood”, “Cbad”, “Cjoint”} are different from one another. Recall that in each of the 10 experiments, the “Cwild” sets are also different from one another. In other words, each of the 10 experiments has their own {“Cgood”, “Cbad”, “Cjoint”, “Cwild”}.

**Lab Face Alignment** Using lab faces to help matching wild faces, first, we need to geometrically normalize them to a common coordinate system. We choose to align lab faces...
Fig. 14. Lab face alignment. (a) alignment template building; (b) wild faces and aligned lab faces.

TABLE VIII
MEAN EERs OF THE 10-FOLD EXPERIMENTS ON WILD FACE PAIR MATCHING WITH FOUR DIFFERENT COHORT CONDITIONS.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Czero</th>
<th>Cgood</th>
<th>Cbad</th>
<th>Cjoint</th>
<th>Cwild</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor</td>
<td>0.3047</td>
<td>0.3020</td>
<td>0.2953</td>
<td>0.2857</td>
<td>0.2537</td>
</tr>
<tr>
<td>LBP</td>
<td>0.3477</td>
<td>0.2917</td>
<td>0.2857</td>
<td>0.2657</td>
<td>0.2570</td>
</tr>
</tbody>
</table>

according to the wild ones. As is known, faces in the FRGC ver2.0 database are frontal still faces. To build an alignment template, we select a set of nearly frontal images from the LFW database and then use a publicly available tool [47] to automatically locate four eye corners on each face image. The four eye corners are then used to determine the two eye centers. Fig. 14 (a) illustrates our approach on building the alignment template. The size of the template is equal to the size of the LFW face images, i.e., $w = h = 250$. The two eye centers of the template are set as the average positions of eye centers of those selected nearly frontal LFW faces. With the two eye centers, we can calculate the distance between them ($d = 46$) as well as the midpoint of them ($[dx, dy] = [125, 113]$). Finally, lab faces from the FRGC ver2.0 database can be geometrically normalized using this built template. In Fig. 14 (b), we show two wild faces and two aligned lab faces.

Experimental Results The mean EERs of the 10-fold experiments are reported in Table VIII. Differing from the result of matching lab faces, for matching wild faces, using cohort samples of good quality achieves the worst performance. With “Cbad” and “Cjoint”, cohort normalization achieves slightly better performance than with good cohort samples. The best performance is obtained by using wild cohort samples. For example, for Gabor and LBP features, using “Cwild” leads to 4.83% and 3.47% lower EERs than using “Cgood”. These results suggest using cohort samples selected from real-world images achieve better performance than using cohort samples obtained under lab environments, when matching wild faces.

D. A Cohort Set with the Maximum or the Minimum Possible Number of Subjects?

To find the answer of the last question proposed in Section I, in this section, we perform a set of experiments on the FRGC ver2.0 database focusing on lab face verification. As shown in the above experiments, for matching lab faces, cohort normalization with good cohort samples can lead to an obvious improvement to the baseline system. Hence, in the following experiments, we use only good cohort samples for cohort normalization. We organize two different cohort sets. The first cohort set contains as many subjects as possible, each subject with the fewest possible samples. While there are the fewest possible subjects, each subject with as many samples as possible, in the second cohort set. We use the same 5 folds as those used in the former experiments on the FRGC ver2.0 database. If a fold is chosen for the cohort use, we select only cohort samples captured under well controlled environments for constituting our two cohort sets. The 6 combinations of target, query and cohort samples are listed in Table IX. Here, we use “Cgood1” to represent the first cohort set and “Cgood2” for the second cohort set. In comparison with the systems in Table IV, only the cohort conditions are changed. The other experimental settings remain fixed. Thus, the verification configuration of the 5 folds remains the same as that listed in Table VI. We describe the cohort configuration of the 5 folds in Table X. For each fold, the numbers of cohort samples of both “Cgood1” and “Cgood2” are the same. In Table X, we list not only the number of subjects in each cohort set, but also the minimum and maximum numbers of samples of each subject.

We summarize the results in Table XI. As observed, for matching either good or bad queries, cohort normalization with “Cgood1” and “Cgood2” leads to similar performance.

VI. Conclusion

In this paper, we introduced picture-specific cohort score normalization to pair matching of faces captured from news articles on the web. With the proposed framework, we showed...
that the cohort information alone has a certain discrimination power which is just marginally worse than the raw matching score. When this information is properly extracted, i.e., using polynomial regression in our case; and is appropriately combined with the raw matching scores, i.e., using logistic regression, we nearly always observed an improvement in the system performance over the baseline system. We validated our approach on the LFW database and achieved competitive performance with the current state of the art. In addition to the above finding, we have also improved our understanding on the behavior of cohort normalization. In particular, a larger cohort set size gives more stable and often better results to a point before the performance saturates and slightly reduces. For the impact of cohort quality, cohort samples with different quality indeed produce different cohort normalization performance. Generally, for matching lab faces, using cohort samples of good quality leads to much better performance than using bad cohort samples. However, for matching wild faces, using cohort samples obtained under well controlled lab environments achieves much worse performance. In contrast, using wild cohort leads to the best performance. Furthermore, the results have shown that the selection of a cohort set, when including only cohort samples with good quality, between the maximum and the minimum possible numbers of subjects has little impact on lab face verification performance. These results can provide researchers with useful suggestions on organizing a suitable cohort/background set for face recognition.

As biometric systems are expected to operate under challenging conditions, cohort normalization certainly offers an option to render them more robust. As presented above, when using background samples, researchers try to exploit not only discriminative information from them but also some other helpful information such as identity and representative information. Thereby, to achieve a thorough understanding of background behavior, much more efforts need to be made by using different background-based approaches.

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. This work has been partially supported by the European Commission COST Action IC1106 and by Biometrics Evaluation and Testing (BEAT, an EU FP7 project with grant no. 284989).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Query</th>
<th>Czero</th>
<th>Cgood1</th>
<th>Cgood2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabor</td>
<td>Qgood</td>
<td>0.1123</td>
<td>0.0607</td>
<td>0.0818</td>
</tr>
<tr>
<td></td>
<td>Qbad</td>
<td>0.2867</td>
<td>0.2240</td>
<td>0.2273</td>
</tr>
<tr>
<td>LBP</td>
<td>Qgood</td>
<td>0.0746</td>
<td>0.0471</td>
<td>0.0489</td>
</tr>
<tr>
<td></td>
<td>Qbad</td>
<td>0.3185</td>
<td>0.2348</td>
<td>0.2337</td>
</tr>
</tbody>
</table>

REFERENCES


