Complementary Cohort Strategy for Multimodal Face Pair Matching

Yunlian Sun, Kamal Nasrollahi, Zhenan Sun, *Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

***Abstract*— Face pair matching is the task of determining whether two face images represent the same person. Due to the limited expressive information embedded in the two face images as well as various sources of facial variations, it becomes a quite difficult problem. Toward the issue of few available images provided to represent each face, we propose to exploit an extra cohort set (identities in the cohort set are different from those being compared) by a series of cohort list comparisons. Useful cohort coefficients are then extracted from both sorted cohort identities and sorted cohort images for complementary information. To augment its robustness to complicated facial variations, we further employ multiple face modalities owing to their complementary value to each other for the face pair matching task. The final decision is made by fusing the extracted cohort coefficients with the direct matching score for all the available face modalities. To investigate the capacity of each individual modality on matching faces, the cohort behavior, and the performance achieved using our complementary cohort strategy, we conduct a set of experiments on two recently collected multimodal face databases. It is shown that using different modalities leads to different face pair matching performance. For each modality, employing our cohort scheme significantly reduces the equal error rate. By applying the proposed multimodal complementary cohort strategy, we achieve the best performance on our face pair matching task.**

***Index Terms*— Face recognition, multimodal fusion, RGB-D, cohort information.**

I. Introduction

THE ANALYSIS of human faces has been a long standing problem in computer vision and pattern recognition. It has received significant attention due to its wide applications in access control and video surveillance (for example, for human identity recognition) [1], human-computer interaction (for example, for emotion analysis) [2] and demography (for example, for gender recognition, ethnicity classification and age estimation) [3]. Among these applications, automatically recognizing humans by analyzing their faces, i.e., face recognition, has been one of the most extensively studied problems in the scientific community. A face recognition system can be an identification system, a verification expert, or a pair matching system. In both identification and verification scenarios, there is a pre-enrolled face database, storing the template for representing each registered user [4]. An identification system aims to decide which subject in the pre-enrolled database, a probe face image comes from, while the task of verification is to determine whether a query face image belongs to the user represented by its claimed template. Differing from these two tasks, in face pair matching, there is no pre-enrolled template database. Given two face images, the goal is to decide whether they are from the same person (a genuine pair) or not (an impostor pair) [5]. Notice that in this task, the only available information is the photometric information embedded in the two images, which makes this task extremely hard. This is the focus of our work.

In 2008, the release of the Labeled Faces in the Wild (LFW) face database makes face pair matching become a popular topic in both the research and industrial community [5]. To well handle diverse facial variations presented on the two face images being compared, a number of powerful facial descriptors have been devised. These facial features are either handcrafted or learned. The patch-based LBP codes [6], the learning-based (LE) descriptor [7] and the discriminant face descriptor (DFD) [8] are some representative facial representations. Very recently, using a large deep neural network to derive an elaborated facial representation has shown a great potential. Deep learning performs well in particular when large training sets are available. It has seen great success in various domains including computer vision, language modeling and speech. Two representative methods are the DeepFace [9] and the DeepID [10]. Instead of developing useful facial representations, another category of approaches aim to learn an appropriate similarity measure to better drive the matching. Logistic discriminant metric learning (LDML) [11] and cosine similarity metric learning (CSML) [12] are among these algorithms.

Besides the face pair matching scenario, in many identification and verification applications, due to the difficulty of gathering face images and the cost for storing and processing them, only very few or even single sample is provided for each identity [13]. In such cases, we do not have enough information to predict the variations in the test samples, either. To address these problems, quite a few recent attempts concentrate on exploiting an additional set of face images

Manuscript received July 5, 2015; revised October 27, 2015 and December 11, 2015; accepted December 17, 2015. Date of publication December 25, 2015; date of current version February 24, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316300, in part by the National Science Foundation of China under Grant 61273272, and in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02080007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Matti Pietikainen. *(Corresponding author: Zhenan Sun.)*

Y. Sun, Z. Sun, and T. Tan are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yunlian.sun@nlpr.ia.ac.cn; znsun@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

K. Nasrollahi is with the Visual Analysis of People Laboratory, Aalborg University, Aalborg 9000, Denmark (e-mail: kn@create.aau.dk).

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.2015.2512561

to compensate for the lack of representative information [6], [14], [15]. Generally, face images in this extra set do not belong to the subject/subjects being compared/tested. Several terms representing this face set in the literature include the background database, cohort set, generic set, library, memory, etc. In this article, we use cohort to indicate this concept, face images in the cohort set are then called cohort samples. Thus, we are interested in exploiting useful information from a set of cohort face images for our face pair matching task.

Parallel to the development of facial feature extraction and face matching [16], different face modalities (depth, thermal, etc) have also been exploited to handle complicated facial variations [17], [18]. Generally, different face modalities behave differently when facing different imaging conditions/degradations, for example, depth images can handle changes caused by different poses to some extent, but susceptible to expression variations. For thermal images, they are sensitive to the temperature changes of the surrounding environment. Thus, a reasonable way to utilize these modalities is to fuse them for the reduction of diverse corrupting factors, which usually affect different modalities in different degrees. As for our face pair matching task, fusion of different modalities certainly offers an alternative to provide supplementary information for the lack of information. Therefore in this work, we focus on multimodal face pair matching. Specifically, we propose a modality-specific Cohort List Comparison (CLC) scheme to perform this task. To the best of our knowledge this work proposes the first multimodal cohort based face pair matching system. More specifically, the main contributions of this work are:

1) Modality-Specific Cohort List Comparison. To perform multimodal face pair matching using cohort, we propose to independently perform a series of cohort list comparisons for each individual modality, including both Cohort Identity List Comparisons (CILC) and Cohort Sample List Comparisons (CSLC). By using both approaches, we expect cohort coefficients extracted by one approach to be complementary to those extracted by the other approach, for our face pair matching task.

2) Application to Multimodal Face Pair Matching. A series of 1-modal, 2-modal and 3-modal face pair matching experiments are conducted on two recently collected multimodal face databases to discover the potential of each individual modality on matching faces, the cohort behavior and the performance of fusing different modalities.

3) Analysis of Cohort List Comparison. We further provide an analysis of our CLC including its differences from several existing cohort investigation approaches and the complementarity of CILC to CSLC for our face pair matching task.

The rest of the paper is organized as follows: Section II gives some existing work on face recognition using cohort information and different modalities. Our proposed multimodal complementary cohort strategy is detailed in Section III. Section IV goes on to test our algorithm on two recently collected multimodal face databases. Section V is devoted to the discussion of several issues involved in the proposed approach. Finally, we conclude the whole work and further give some interesting future work in Section VI.

II. Related Work

In this section, we provide a short literature review on face recognition using cohort information and different modalities.

A. Face Recognition Using Cohort Information

Cohort samples, whose identity is different from those of samples being compared, are early used to improve the recognition performance of a biometric system. It was initially proposed for speaker recognition [19], [20], and then successfully applied to fingerprint verification [21]–[23], face verification [24], sparse representation based face identification [25], unconstrained face pair matching [14], [26] and multi-biometrics [27]. The interested reader is referred to [14] for further information on using cohort to improve a biometric system. Here, we present only several existing cohort/cohort similar techniques related to face recognition.

For comparing two faces under significantly different settings, Schroff et al. proposed to describe an input face image by an ordered list of identities from a Library [15]. In the ordered list, identities are ranked according to their similarity to the input face image. The similarity between two face images is then computed as the similarity of their corresponding ordered lists. For the same purpose, Yin et al. proposed to "associate" a test face with alike identities from an extra generic identity data set [28]. With the associated faces, the likelihood whether two input faces are from the same person or not can then be discriminatively "predicted". To apply the traditional sparse representation-based classifier [29] to undersampled face identification, an auxiliary intra-class variant dictionary was employed in [30] to represent possible variations between training and test images. The dictionary atoms, representing intra-class sample differences, were computed from a set of generic faces. To address the same problem, in [31], a sparse variation dictionary was learned from a generic set to improve the test sample representation by a single training sample per person. Liao et al. [32] 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 faces together with one of the two input faces. To handle unconstrained face pair matching, Tistarelli et al. developed a picture-specific cohort score normalization approach [14], by extracting discriminative cohort coefficients from a pool of sorted cohort samples using polynomial regression. Wolf et al. learned a discriminative model exclusive to the two face images being compared from a set of background samples [6]. In another work, an additional identity data set was employed for building a set of either attribute or simile classifiers [33]. Li et al. trained a Gaussian Mixture Model (GMM) on the spatial-appearance features by employing an independent training set [34], each Gaussian component was then used to build correspondence of a pair of features to be matched between two face images. A similar GMM with diagonal covariances was trained on dense patch features in [35] to compute the fisher vector representation of a particular face image. A training set, which does not include samples of the identity/identities being compared, was employed.

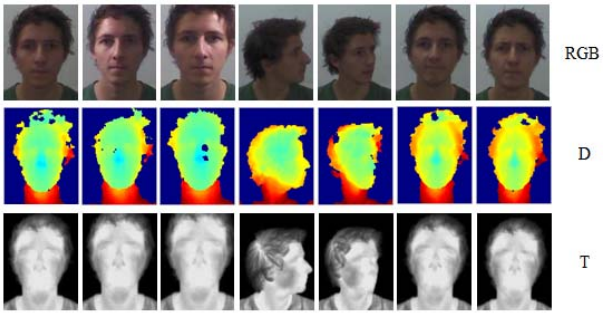


Fig. 1. Several face images of RGB, depth and thermal modalities from the database of [43].

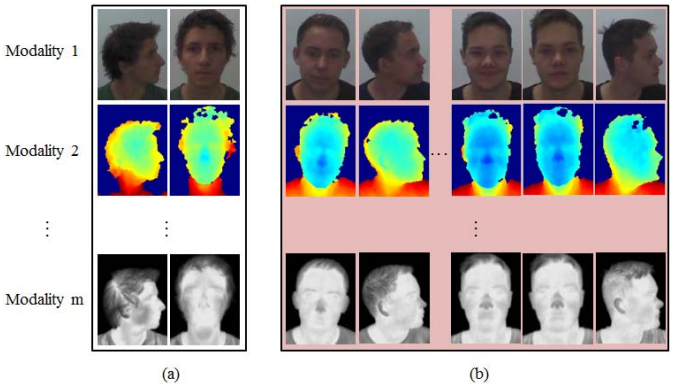


Fig. 2. Examples of test and cohort face images used in our framework. (a) Two test face images; (b) cohort face images.

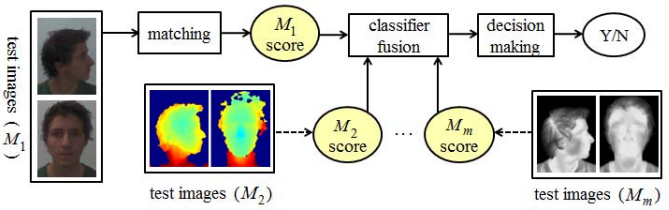


Fig. 3. Framework of multimodal face pair matching on the score level fusion. $M_1, M_2, \ldots, M_m$ represent the m modalities.

B. Face Recognition Using Different Modalities

In the literature, some significant attempts have been devoted to explore the usefulness of different face modalities. 3D face recognition is among these techniques [17]. With a 3D model, face images with different poses can be aligned owing to the well captured facial geometry information. Furthermore, the 3D face shape is illumination invariant. This technique does offer a more suitable description of facial features than 2D models, increasing the robustness to viewpoint and lighting variations. However, the low acquisition efficiency and high cost of 3D scanners limit its use in practical applications. With progress in sensor technology, low cost sensors have been developed capable of capturing less accurate 3D information in the form of RGB-D images. The Kinect is among such devices, which can provide synchronized images of both color (RGB) and depth (D). The color image depicts the appearance and texture information of a face, while the depth map measures the distance of each pixel from the camera. Exploiting RGB-D images has become more and more popular in tackling various computer vision problems [36]–[38]. In [39]–[42], interesting work on using RGB-D images for face recognition was presented. Another less commonly used modality is the infrared imagery [18]. A thermal (T) infrared image records the amount of infrared radiation emitted by an object. The amount of radiation increases with temperature, therefore, this imagery allows us to see variations in temperature. When viewed through a thermal imaging camera, humans and other warm-blooded animals become easily visible against the cool environment, with or without visible illumination. In Fig. 1, we show several RGB, depth and thermal images of one person from the face database of [43].

III. Multimodal Complementary Cohort Strategy

In this work, we concentrate on multimodal face pair matching. Consequently, in our problem, for each face of a pair, we have multiple synchronized images corresponding to different modalities, as illustrated in Fig. 2 (a). Suppose m shows the number of the available modalities. For example, in RGB-D based face recognition, $m = 2$. For each particular modality, after matching the corresponding two face images, we can get a similarity score. In this work, we use cosine similarity as the similarity measure. Euclidean and Hellinger distances, however, can be equally used in our proposed framework. In order to both effectively and efficiently utilize information provided by different modalities, we employ the score level fusion strategy, following the taxonomy in information fusion [44]. In comparison to the decision level fusion, score level fusion preserves more information, while requires much lower complexity than the feature level fusion. The schematic process is depicted in Fig. 3.

To better drive the matching, we exploit background information from an extra cohort set. Similarly, in the cohort set, for each cohort face, we incorporate synchronized images from different modalities, as shown in Fig. 2 (b). Our proposed modality-specific approach for multimodal face pair matching using cohort information is represented in Fig. 4. Take the RGB modality as an example, we have two input RGB face images together with a set of cohort RGB face images. By ranking all the cohort images according to their similarity, namely cosine similarity, to the two input face images, we can get two ordered cohort lists, respectively. Next, cohort information/coefficients embedded in the two sorted cohort lists can be extracted and further combined with the direct matching score of the two test face images, i.e., their cosine similarity denoted as $rawSC$, forming the final RGB contribution ($M_1 contri$). In the same way, we can obtain contributions of other modalities ($M_2 contri, \ldots, M_m contri$). The final decision is made by fusing all the m contributions.

The method we developed for extracting cohort information is based on a series of cohort list comparisons. In CLC, we include both cohort identity list comparisons and cohort sample list comparisons. In both comparisons, we have the same cohort face images from a set of subjects. We call

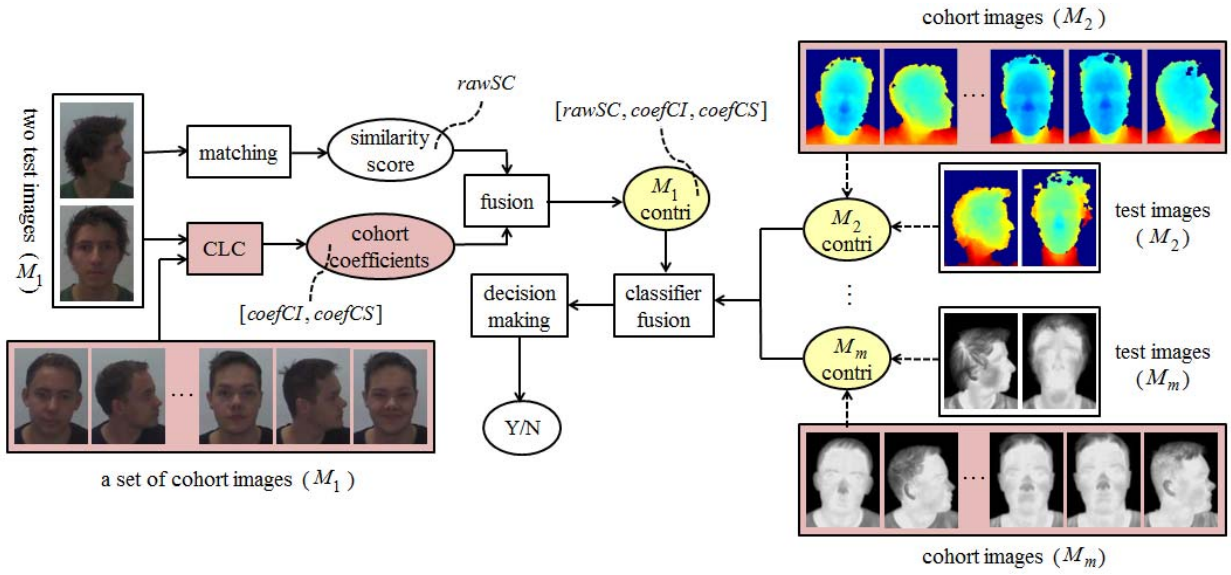


Fig. 4. Framework of the proposed multimodal complementary cohort strategy for face pair matching. $M_1contri, M_2contri, \ldots, M_mcontri$ represent contributions of the m different modalities to the final decision matching.

Fig. 5. Visual interpretation of CILC. The cohort set used to compute the ranked cohort list contains 1,700 images from 17 cohort subjects, each with 100 images. Therefore, each ranked cohort list denotes an ordering of all the 1,700 images. The figure displays only the top ten positions, corresponding to the first 10 closest cohort samples to the test face image. Cohort images marked by the same kind of squares (solid or dashed) share the same cohort identity. Face images are from the database of [43].

these subjects cohort subjects or cohort identities. In CILC, cohort information is extracted from sorted cohort identities, using their positions in the ranked cohort list. However, we investigate cohort information from sorted cohort face images in CSLC by means of ranked cohort scores, i.e., similarity scores between cohort samples and test face images. The reason why we use these two different cohort algorithms is that we expect cohort information discovered by them to be supplementary to each other for our face pair matching task (This will be shown in the experimental result section). Let I_1 and I_2 represent the two test face images, say in the RGB modality, being compared. The cohort set C is composed of H face images from N subjects. We denote the two sorted cohort lists, obtained by sorting cohort samples in C according to their similarity to I_1 and I_2, as $C_1 = [c_{11}, \ldots, c_{1H}]$ and $C_2 = [c_{21}, \ldots, c_{2H}]$, where $c_{11}(c_{21})$ is the nearest cohort sample to $I_1(I_2)$ and $c_{1H}(c_{2H})$ is the furthest one. Next, we explain the details of how to implement CILC and CSLC using C_1 and C_2.

A. Cohort Identity List Comparison

A visual explanation about how CILC works is displayed in Fig. 5. We have one genuine pair and one impostor pair. For the genuine one, the two images faceA1 and faceA2 are captured with different poses, whereas faceA1 and faceB1 of the impostor pair are both frontal faces. By ranking all the cohort images with respect to their closeness to the three test images, we can get three corresponding ranked cohort lists.

Our CILC approach draws its motivation from the observation that if two people look like each other, then they should to some extent share similar expressions, profile views, etc. Let us look at faceA1 and its cohort list in Fig. 5. Based on the frontal view, the two cohort subjects marked by solid and dashed squares look similar to the subject pictured by faceA1. Correspondingly, they should have similar profile views to the subject of faceA1. This is verified in Fig. 5 by faceA2 and its cohort list, where the side views of the two cohort subjects locate at positions close to faceA2, which is the profile view of the subject represented by faceA1. Furthermore, the positions of their side and frontal views in the two cohort lists are not far from each other. For example, for the cohort subject marked by the solid square, its side views locate at positions $[2, 5]$ in the first cohort list, while the positions of frontal views in the second cohort list are $[1, 7, 8, 9, 10]$. For an ordering of 1,700 images, $[2, 5]$ and $[1, 7, 8, 9, 10]$ are close positions. On the other hand, if two people look quite different from each other, their top ranked cohort lists should include quite few images of common cohort subjects even when they are captured under similar imaging conditions, as substantiated in Fig. 5 that faceA1 and faceB1 share zero common cohort identities. Based on this observation, we can describe a test face image by its ranked cohort list. For comparing two test face images, we calculate the similarity of their corresponding ranked cohort lists as cohort information to assist the comparison.

Our proposed CILC is similar to the Doppelgänger list comparison developed in [15]. However, each cohort identity appears only once in the Doppelgänger list, while in our cohort set, each cohort identity can have multiple images, thus appearing multiple times and corresponding to multiple positions in the ranked cohort list. As discussed in [15], for distinguishing between genuine and impostor face pairs, top positions in a ranked cohort list include far more discriminative information than later ones. Accordingly, we employ only the first K cohort samples in C_1 and C_2 for cohort coefficient computation, i.e., $Coh_1 = [c_{11}, \ldots, c_{1K}]$ and $Coh_2 = [c_{21}, \ldots, c_{2K}]$. The developed algorithm for computing cohort coefficients embedded in the ranked cohort lists is described in Alg. 1.

We use $coefCI$ to represent the cohort coefficients extracted by CILC, namely the similarity of I_1 and I_2 determined by the similarity of their ranked cohort lists rather than their cosine similarity. We include three similarity in $coefCI$, which are computed in three different levels. When computing $sim1$, for each cohort identity, we employ only its closest sample to the test image. That is only its first cohort position is considered. For each cohort identity, let s and t represent the numbers of its cohort positions previous to K in the two sorted cohort lists, C_1 and C_2, respectively. During the computation of $sim2$, we consider the first r cohort positions, r is the minimum of $\{s, t\}$. When computing $sim3$, all the cohort positions of a cohort subject are considered as long as they are previous to K. The computation of $sim1$, $sim2$ and $sim3$ is based on the weighted voting scheme of neighbors proposed by Jarvis et al. for clustering [45]. In our following

Algorithm 1 Cohort Coefficient Computation by CILC

Input: Coh_1, Coh_2, N, K;
Output: $coefCI$;
Set $sim1 = 0$, $sim2 = 0$, $sim3 = 0$;
Set $ct1 = 0$, $ct2 = 0$, $ct3 = 0$;
for $i = 1$ to N **do**
  Let $pos_1 = [r_{11}, \ldots, r_{1s}]$ and $pos_2 = [r_{21}, \ldots, r_{2t}]$ represent the ranks of the i^{th} cohort subject's images in Coh_1 and Coh_2;
  $r = min(s, t)$;
  if $r \geq 1$ **then**
    $sim1 = sim1 + \frac{(K+1-r_{11})\times(K+1-r_{21})}{K\times K}$;
    $ct1 = ct1 + 1$;
    for $j = 1$ to r **do**
      $sim2 = sim2 + \frac{(K+1-r_{1j})\times(K+1-r_{2j})}{K\times K}$;
      $ct2 = ct2 + 1$;
    end for
    for $p = 1$ to s **do**
      Find the closest value in pos_2 to r_{1p}, denoted as r_{2q};
      $sim3 = sim3 + \frac{(K+1-r_{1p})\times(K+1-r_{2q})}{K\times K}$;
      $ct3 = ct3 + 1$;
    end for
    for $q = 1$ to t **do**
      Find the closest value in pos_1 to r_{2q}, denoted as r_{1p};
      $sim3 = sim3 + \frac{(K+1-r_{1p})\times(K+1-r_{2q})}{K\times K}$;
      $ct3 = ct3 + 1$;
    end for
  end if
end for
$coefCI = [\frac{sim1}{ct1}, \frac{sim2}{ct2}, \frac{sim3}{ct3}]$;

accumulated similarity,

$$sim = sim + \frac{(K + 1 - r_{1u}) \times (K + 1 - r_{2v})}{K \times K} \tag{1}$$

r_{1u} and r_{2v} are positions of the two cohort samples from one cohort subject in the two ordered cohort lists, respectively. Then, if both cohort samples are the closest ones to their corresponding test faces, namely $r_{1u} = 1$ and $r_{2v} = 1$, their contribution to the similarity computation can be calculated as $K \times K$. However, if both locate at the furthest positions, i.e., $r_{1u} = K$ and $r_{2v} = K$, their contribution turns to 1. This similarity calculation scheme is in keeping with the notion that the more similar a cohort sample to the test sample, the more information about the local density of the test image that cohort sample can provide. By dividing $K \times K$, we aim to get normalized contribution. A further normalization procedure is followed by dividing the number of accumulated contributions, i.e., $ct1$, $ct2$ and $ct3$ in Alg. 1. The extracted $coefCI$ to some degree provides invariance to the direct matching score $rawSC$ across different expressions, poses, etc. A detailed discussion about the usefulness of $coefCI$ in matching faces will be given in Section V. In the following sections, for simplicity, we use $coefCI = [sim1, sim2, sim3]$ to represent $coefCI = [\frac{sim1}{ct1}, \frac{sim2}{ct2}, \frac{sim3}{ct3}]$.

B. Cohort Sample List Comparison

Note that in the above described CILC, cohort coefficients are calculated by positions of each cohort identity in the ranked cohort lists. Now we keep on extracting cohort information embedded in them, but by means of sorted cohort scores between cohort samples and test samples. To do so, we expect the extracted cohort coefficients will provide some complementary information to those discovered by CILC. Given I_1 and I_2 and their ranked cohort lists $C_1 = [c_{11}, \ldots, c_{1h}, \ldots, c_{1H}]$ and $C_2 = [c_{21}, \ldots, c_{2h}, \ldots, c_{2H}]$, we employ the picture-specific cohort ordering strategy developed in [14] for CSLC. For comparing two face pictures, this technique stems from the observation that cohort samples, sorted by their closeness to the reciprocal face picture, produce some discriminative information between genuine and impostor pairs. Polynomial regression is then used to extract this discriminative information.

Let $sc_1 = [sc_{11}, \ldots, sc_{1h}, \ldots, sc_{1H}]$ denote the cohort score list of I_1 and all the cohort samples in C_2. That is, sc_{1h} is the similarity score between I_1 and c_{2h}. Similarly, $sc_2 = [sc_{21}, \ldots, sc_{2h}, \ldots, sc_{2H}]$ lists cohort scores of I_2 and all the cohort samples in C_1. sc_1 and sc_2 are the two so-called picture-specific cohort score lists. Be warned that the ordering of the cohort score profile for I_1 is determined by I_2; and that of I_2 is determined by I_1. Next, we consider cohort scores in sc_1 and sc_2 as discrete points on two functions of rank orders as follows:

$$sc_{1h} = f_1(h) \tag{2}$$

$$sc_{2h} = f_2(h) \tag{3}$$

where $h = 1, 2, \ldots, H$. Now let us move on to the conception behind the picture-specific cohort ordering strategy in [14]. If I_1 and I_2 are from the same subject, their ranked cohort lists C_1 and C_2 should to some degree look similar. Note that scores in sc_1 are cohort scores between I_1 and all the cohort samples in C_2, which are previously sorted according to their closeness to I_2. Consequently, sc_1 (or f_1) should follow a decreasing profile as the cohort sample order h increases. However, if I_1 and I_2 are from an impostor pair, sc_1 (or f_1) should correspond to a disorganized/flat one. Likewise, we can get a similar conclusion for sc_2 (or f_2).

Now we focus on how to extract this discriminative information between cohort score profiles of genuine and impostor pairs. The two functions are approximated using polynomial regression as follows:

$$f_1(h) \approx w_{1n}h^n + w_{1,n-1}h^{n-1} + \ldots + w_{11}h + w_{10} \tag{4}$$

$$f_2(h) \approx w_{2n}h^n + w_{2,n-1}h^{n-1} + \ldots + w_{21}h + w_{20} \tag{5}$$

where $w_1 = [w_{10}, w_{11}, \ldots, w_{1n}]$ and $w_2 = [w_{20}, w_{21}, \ldots, w_{2n}]$ are the two approximated polynomial coefficient vectors. Further, cohort scores in sc_1 and sc_2 can be approximated by the $n+1$ coefficients in w_1 and w_2, respectively. Finally, we can use w_1 and w_2 to approximately represent the discriminative information included in sorted cohort scores and have $coefCS = [w_1, w_2]$.

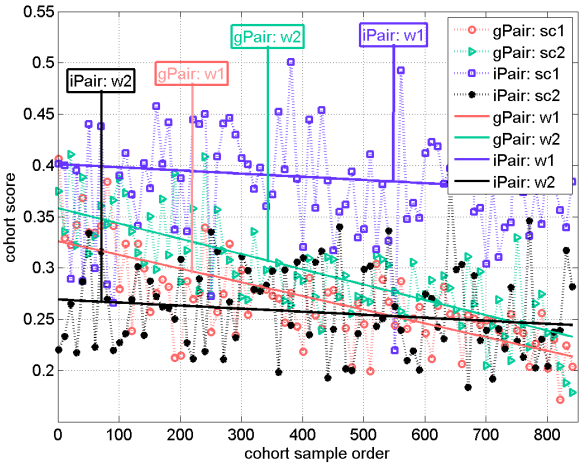


Fig. 6. Cohort score profiles and the corresponding fitting lines for a genuine pair and an impostor pair, computed from the RGB modality of the database of [43]. The cohort set contains 1,700 face images, thus we have a total of 1,700 cohort orders and 1,700 cohort scores in sc_1 and sc_2. However, we employ only half of them for polynomial regression by sampling the 1,700 sorted cohort scores in a step of 2. Accordingly, the cohort orders are from 1 to 850. "gPair" and "iPair" stand for the genuine and impostor pairs, respectively.

In Fig. 6, we display the two picture-specific cohort score profiles (sc_1 and sc_2) as well as their fitted curves (w_1 and w_2) for a genuine pair (denoted as "gPair") and an impostor pair ("iPair") to demonstrate the effectiveness of our extracted $coefCS$. The two pairs are selected from the RGB modality of the database of [43]. We simply employ linear functions to fit cohort score profiles, i.e., the polynomial degree $n = 1$. As can be seen, noises presented on the cohort score profiles are significantly reduced after polynomial regression, making the discriminative information embedded in sc_1 and sc_2 between genuine and impostor pairs more significant. This is illustrated in the fitted lines in Fig. 6, where two lines of the genuine pair follow a downslope path as the cohort sample order increases, whereas the course of the impostor lines is flatter.

C. Classification

Let $rawSC$ signify the matching score of I_1 and I_2 obtained by directly comparing them. By the above presented CILC and CSLC, we can obtain two cohort coefficients $coefCI$ and $coefCS$. Each of $\{rawSC, coefCI, coefCS\}$ contains different but complementary information which can be combined to enhance the classification performance. Up to now, we have finished comparing two test face images by our modality-specific cohort list comparison using the RGB modality. We use $M_1contri$ to indicate the fusion of $[rawSC, coefCI, coefCS]$. By applying the above described procedures to other modalities, we can get their corresponding contributions: $M_2contri$,..., $M_mcontri$. Next, we aggregate these contributions by training a logistic regression classifier [46], which can provide discriminative weights on each parameter of $[M_1contri, M_2contri, \ldots, M_mcontri]$. The final matching score is approximated as:

$$finalSC = P(G \mid M_1contri, M_2contri, \ldots, M_mcontri) \tag{6}$$

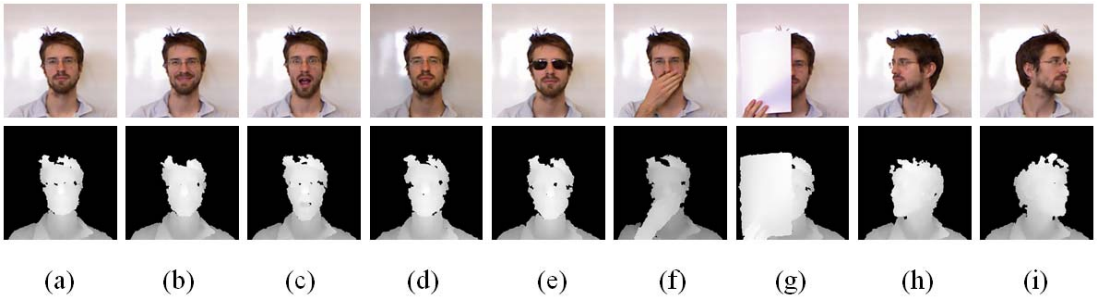


Fig. 7. Face images of one subject from one session corresponding to different facial variations on the KinectFaceDB database. The lower depth maps are aligned with the upper RGB images. (a) Neutral; (b) smiling; (c) mouth open; (d) strong illumination; (e) sunglass occlusion; (f) hand occlusion; (g) paper occlusion; (h) right profile; (i) left profile.

where $P(G \mid M_1 contri, M_2 contri, \ldots, M_m contri)$ represents the probability of being a genuine pair. In other words, the larger $finalSC$ is, the more probable I_1 and I_2 come from the same person.

IV. Application to Multimodal Face Pair Matching

In this section, we apply our proposed modality-specific cohort list comparison to multimodal face pair matching. First, we describe the employed databases of KinectFaceDB [39] and RGB-D-T [43] which are used to determine the actual performance of the proposed approach.

A. The KinectFaceDB Database

KinectFaceDB is a publicly available face database collected by the Kinect sensor [39]. It consists of different face modalities, including the RGB image, depth map and computed 3D point cloud. For each modality, there are 936 shots from 52 individuals. The database was recorded in two different sessions, with 5-14 day intervals between them. In each session, 9 facial variations were recorded, i.e., neutral face, smiling, mouth open, strong illumination, occlusion by sunglasses, occlusion by hand, occlusion by paper, right face profile and left face profile. Thus, each subject has 18 face images for each modality. For the depth map, the authors provide a .bmp depth image and a .txt file with the depth information of each pixel in the original coordinates. We use the .bmp depth map for our experiments. Fig. 7 illustrates the RGB and depth images of one subject from one session. With the RGB and depth images, the 3D coordinates can be computed directly. However, the low quality/depth resolution of the Kinect 3D model makes the face recognition performance not promising, as shown in [39]. Thus, in this work, we use only the RGB and depth modalities for our multimodal face recognition.

The evaluation protocol designed by authors of the KinectFaceDB database is for face identification and face verification. To benchmark our algorithm, we define a new protocol specific for face pair matching. For each modality, we divide the 52 subjects into three folds, which includes 17, 17 and 18 subjects, respectively. With such a division scheme, the subjects are disjoint from one another in the three folds. Consequently, there are $17 \times 18 = 306$, $17 \times 18 = 306$ and $18 \times 18 = 324$ images in the three folds, as listed in Table I. We perform 3-fold cross validation experiments. In each one of the three experiments, one fold is used for evaluation, one is the development set, and the remaining one is used as the cohort set. When we perform face matching in the evaluation and development sets, each face image is compared against all the remaining images. For example, if Fold 1 is used as the development/evaluation set, then the number of total matching is $C^2_{306} = 46,665$, including $17 \times C^2_{18} = 2,601$ genuine and $44,064$ impostor matches. Be warned that, in such defined protocol, we can cover a number of challenging matches. For example, a right profile of a person is compared with his/her occluded face by the paper. If one fold is used as the cohort set, then for both CILC and CSLC, all the images in this fold are used as cohort samples. The data configuration of the 3 folds is listed in Table I.

TABLE I
Data Configuration of the 3 Folds on KinectFaceDB

Fold	1	2	3
# subjects	17	17	18
# images	306	306	324
# total matches	46,665	46,665	52,326
# genuine matches	2,601	2,601	2,754
# impostor matches	44,064	44,064	49,572
# cohort samples	306	306	324

B. The RGB-D-T Face Database

In [43], the authors organized a face database of 51 persons including 45,900 facial images of synchronized RGB, depth and thermal modalities. The Microsoft Kinect for Windows was used to capture RGB and depth images, while thermal images are obtained by the thermal camera AXIS Q1922. This database incorporates three capturing scenarios, recording facial appearance variations due to different poses, expressions and illumination conditions. In each scenario, there are 300 images for each person, with 100 RGB, 100 depth and 100 thermal synchronized pictures. Together with the database, the ground-truth data representing coordinates of the face bounding box is also provided for each image. With such a ground-truth data, face region can be easily detected. Fig. 1 displays several detected face regions from this database.

TABLE II
DATA CONFIGURATION OF INDIVIDUAL FOLDS ON RGB-D-T

Fold	# subjects	# images	# total matches	# genuine matches	# impostor matches	# cohort samples
1/2/3	17	1,700	1,444,150	84,150	1,360,000	1,700

As presented above, each subject has 900 synchronized images from RGB, D and T modalities, with each modality containing 300 images. In our experiments, for each modality, we select only 100 images from the entire 300 images of one subject due to the high similarity between neighbouring images. These 100 images include images from all the three sessions. Therefore, in our protocol, we employ only $51 \times 100 = 5,100$ images for each modality. For each modality, the total 5,100 images are separated into three folds, with each fold 1,700 images from 17 subjects. Similarly, in these three folds, the subjects are disjoint from one another. Next, we can conduct 3-fold cross validation experiments as on the KinectFaceDB database. When we perform face matching in the evaluation and development sets, each face image is compared against all the remaining $1,700-1 = 1,699$ images. In total, we have $C^2_{1,700} = 1,444,150$ matches, including $17 \times C^2_{100} = 84,150$ genuine and $1,360,000$ impostor matches. For the cohort set, we employ all the 1,700 images for both CILC and CSLC. Table II lists the data configuration of individual folds on the RGB-D-T database.

C. Face Pair Matching Pipeline

Here, we provide some details about the preprocessing, feature extraction and face matching involved in our face pair matching pipeline.

1) Preprocessing: When dealing with RGB images, we first convert them to grayscale ones on both databases. For the RGB-D-T database, we directly detect the face region using the available ground-truth information. For KinectFaceDB, based on the manual landmarks provided by the authors, the images are first cropped from 256×256 to 140×120 pixels, with distance between the two eye centers set to 60 pixels. After cropping, the coordinate of the eye center $(vertical, horizontal)$ is equal to $(40, 60)$. For the right (left) profile images, the left (right) eye and the nose are used to align them. These profile images are also cropped to 140×120 pixels. For the right profile images, the horizontal axis of the nose is set to 11 and the vertical axis of the left eye is set to 40. Similarly, for the left profile images, the horizontal axis of the nose is set to 110 and the vertical axis of the right eye is set to 40. In Fig. 8, we show the aligned grayscale images and depth maps. As can be seen, the original depth maps are relatively noisy. For example, the depth values on some pixels are sensed as 0mm but their true values are not zero. To fill these holes, closing operation is further applied to depth maps. The improved depth maps are also illustrated in Fig. 8.

2) Feature Extraction: We compute both Local Binary Patterns (LBP) [47] and Histograms of Oriented Gradients (HOG) [48] for facial feature representation. For computing LBP, we first resize the preprocessed face images to a fixed size 130×100 and then divide each face image into non-overlapping blocks with size 10×10. For each block, we extract a 59-bin uniform LBP histogram. By concatenating histograms of all the blocks, we can get a final feature vector of 7,670 dimension. Before extracting the HOG feature, the preprocessed images are first resized to 128×96. Next, we use the settings adopted in [48] for our feature extraction. The cell size is 8×8, and each block contains 2×2 cells. The number of orientation bins in the histogram of a cell is set to 9. The resulting HOG feature vector is of length 5,940. Since in this work, the focus is on using cohort information and multiple modalities for face pair matching, we do not perform further descriptor fusion, which results in a single descriptor of length 7,670 + 5,940, but simply use LBP and HOG separately.

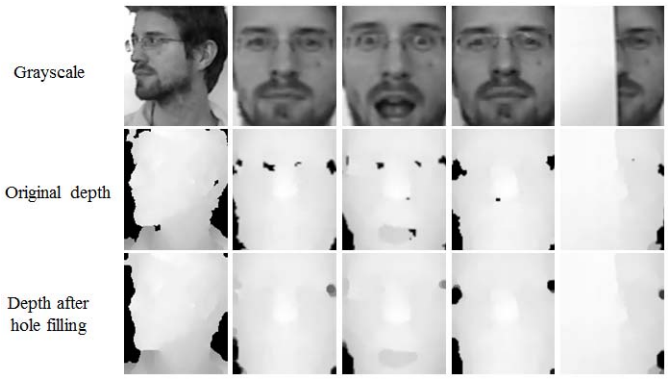


Fig. 8. Aligned grayscale images and depth maps on the KinectFaceDB database.

3) Face Matching: It is worth emphasizing again that in our work, for directly matching two test face images, ranking all the cohort samples according to their similarity to the test sample and computing cohort scores between cohort samples and test samples, we employ the cosine similarity as the similarity measure. For polynomial regression involved in CSLC, we adopt a linear function to fit the two cohort score functions $f_1(h)$ and $f_2(h)$. For KinectFaceDB, there are $H = 306/324$ cohort scores in sc_1 and sc_2, while $H = 1,700$ for the RGB-D-T database. Be warned however, for RGB-D-T, we employ only half of them for polynomial regression by sampling the H sorted cohort scores in a step of 2. That is, we use $\frac{H}{2} = 850$ cohort samples to regress the two cohort score profiles. By doing this, we can perform polynomial regression in a more efficient way but with slight information loss. For training the logistic regression classifier, we use l_2-penalized logistic regression which leads to maximum likelihood estimate [49], [50]. Finally, the Equal Error Rate (EER) is used as the performance evaluation measure [51].

D. Experimental Results

We perform a series of experiments on multimodal face pair matching to investigate the power of each particular modality

TABLE III

EERs (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY USING TWO INDIVIDUAL MODALITIES ON KINECTFACEDB

Feature	LBP		HOG	
Modality	RGB	D	RGB	D
Fold1	**42.37**	48.06	**43.87**	46.14
Fold2	**44.29**	47.33	**46.17**	46.33
Fold3	**42.48**	47.46	**44.66**	45.86
Mean	**43.05**	47.62	**44.90**	46.11

TABLE IV

EERs (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY USING THREE INDIVIDUAL MODALITIES ON RGB-D-T

Feature	LBP			HOG		
Modality	RGB	D	T	RGB	D	T
Fold1	**27.04**	29.79	31.42	29.06	31.15	**27.87**
Fold2	**25.18**	30.44	29.27	29.87	30.82	**29.27**
Fold3	**27.95**	34.95	34.01	**33.48**	34.81	35.22
Mean	**26.72**	31.73	31.57	30.80	32.26	**30.79**

on matching faces, the cohort behavior and the matching performance achieved by fusing different modalities.

1) Potential of Individual Modality: The first sequence of experiments is conducted to find out the capacity of each individual modality on matching faces. In order to get a clear insight of this, in these experiments, we directly use the raw matching score $rawSC$ for the classification instead of getting help from cohort. We report the results of the 3-fold experiments on the two databases as well as their mean EERs in Table III and Table IV, respectively. From Table III, it is easy to conclude that with both LBP and HOG, using the RGB modality achieves much lower EERs than using depth maps on the KinectFaceDB database. On the RGB-D-T database, when LBP is used as the facial feature, the best performance is obtained by the RGB modality. Depth and thermal modalities lead to similar performance. However, when using HOG as the feature, we get the best performance with the thermal modality in most cases. As shown in the above results, LBP works better on RGB images compared to the other modalities because LBP is known to be more dependent on the texture [52], which is more visible in RGB than in thermal and depth. HOG, however, is more dependent on the edge information [48], [53], [54]. The edges are more pronounced in thermal compared to RGB and depth. From Fig. 1, we can observe that edge information in depth is the least noticeable. This is consistent with the above experimental resutls that HOG works better with thermal and RGB than with depth.

2) Impact of Parameter K on CILC Performance: As described in Section III-A, for all the H positions in the ranked cohort list, we employ only the top K positions. Therefore, it is interesting to find out the impact of different K values on the generalization performance. For this issue, we still perform experiments independently for each modality. In addition, we do not take cohort coefficients determined by CSLC into account, as it does not change the influence of K. Finally, the input of our classifier becomes $[rawSC, coefCI]$. In Table V, we show the EERs obtained by CILC with different values of K on one fold of KinectFaceDB, using the LBP feature.

TABLE V

COMPARATIVE EERs (%) OBTAINED BY CILC WITH DIFFERENT K VALUES FOR SINGLE MODALITY ON KINECTFACEDB, USING LBP

Modality	noCLC	K=50	K=100	K=200	K=300
RGB	42.37	32.06	34.45	32.87	34.18
D	48.06	41.41	44.18	43.21	44.83

TABLE VI

COMPARATIVE EERs (%) OBTAINED BY CILC WITH DIFFERENT K VALUES FOR SINGLE MODALITY ON RGB-D-T, USING LBP

Modality	noCLC	K=100	K=300	K=500	K=700
RGB	27.04	24.20	24.42	23.81	23.77
D	29.79	30.38	27.48	26.55	26.52
T	31.42	28.71	30.15	30.17	29.34

The results of RGB-D-T are reported in Table VI. We further list the EERs achieved by using only the raw matching score $rawSC$, denoted as "noCLC" in the two tables. It is observed that K does not affect the performance significantly. In all our following experiments, therefore we simply choose $K = 50$ and $K = 300$ for KinectFaceDB and RGB-D-T, respectively.

3) Discriminative Information Discovered by CSLC: Now we visualize the discriminative information embedded in sorted cohort scores discovered by CSLC for each modality. The experiments are conducted on one fold of the RGB-D-T database. Thus, we have 84,150 genuine and 1,360,000 impostor pairs. For each pair, we can get two picture-specific cohort score profiles sc_1 and sc_2, each of which is a single vector of 850. Correspondingly, we can get a total of $84,150 \times 2 = 168,300$ genuine and $1,360,000 \times 2 = 2,720,000$ impostor cohort score profiles. Next, we respectively compute the mean and standard deviation of these cohort score profiles. Fig. 9 shows the cohort score distribution using LBP for each modality. By illustrating the mean of large numbers of cohort score profiles, we can get smoother cohort score distribution than directly displaying single cohort score profile, i.e., the noisy profiles shown in Fig. 6. As noticed from these figures, the discriminative information between genuine and impostor pairs is made explicit by CSLC.

4) Modality-Specific Cohort Behavior: Now we perform a group of experiments to unearth the cohort behavior. Recall that for each modality, the final contribution to matching the two test face images is $[rawSC, coefCI, coefCS] = [rawSC, sim1, sim2, sim3, w1, w2]$. By cohort behavior, we are driving at the amount of useful information that each individual cohort coefficient can offer to the pair matching task, namely their contributions. We quantitatively analyze this cohort information by computing how much improvement we can achieve in the presence of different cohort coefficients compared to the baseline system using only $rawSC$. We use a group of tags to represent different systems, as listed in Table VII. Thus, "noCLC" denotes the baseline system using only $rawSC$ and "CILC1" is the system using $[rawSC, sim1]$. "CILC" indicates integrating only cohort identity list comparison with the direct matching score, i.e.,

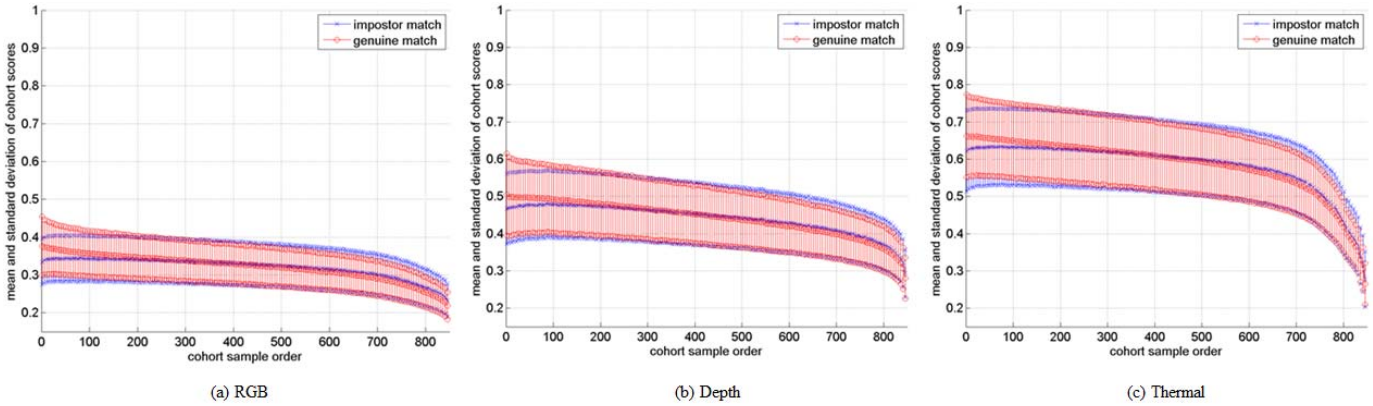


Fig. 9. Distribution of cohort scores generated by ordered cohort samples for the three modalities on the RGB-D-T database, using LBP.

TABLE VII
TAGS USED TO REPRESENT DIFFERENT SYSTEMS

Tag	noCLC	CILC1	CILC2	CILC3	CILC
System	$[rawSC]$	$[rawSC, sim1]$	$[rawSC, sim2]$	$[rawSC, sim3]$	$[rawSC, coefCI]$
Tag	CSLC1	CSLC2	CSLC	CLC	
System	$[rawSC, w1]$	$[rawSC, w2]$	$[rawSC, coefCS]$	$[rawSC, coefCI, coefCS]$	

$[rawSC, coefCI]$. In the system of "CLC", we implement our proposed CLC by including both CILC and CSLC.

The mean EERs of the 3-fold cross validation experiments using both LBP and HOG are reported in Table VIII and Table IX, for KinectFaceDB and RGB-D-T, respectively. As observed, by using either CILC or CSLC, the EER is significantly reduced on both databases. However, CSLC achieves much lower EER than CILC. Take the HOG feature as an example, on RGB-D-T, the reduced EERs by CILC are 5.61%, 3.98% and 4.97% for RGB, depth and thermal modalities, respectively. While the three figures achieved by CSLC are 13.99%, 10.14% and 14.89%, respectively. One reason might be the largely suppressed noise on sorted cohort score profiles by polynomial regression as shown in Fig. 6. Another reason might be the small number of cohort identities included in the cohort set. In our RGB-D-T experiments, the cohort set contains 1,700 face images from only 17 cohort identities, while in [15], the authors employed a Library of 750,000 face images from 337 subjects. By integrating CILC into CSLC, i.e., our proposed CLC, we do get some improvement. In Section V, we shall discuss the necessity of CILC, by showing its complementarity to CSLC for our face pair matching. Further, we observe that the three similarity measures $[sim1, sim2, sim3]$ we designed lead to different performance, by combining them together, we achieve the best performance in most cases. Similarly, in some cases, w_1 and w_2 discovered by CSLC produce largely different performance, however by fusing them, we get better results than using either of them. This demonstrates again the effectiveness of the picture-specific cohort ordering strategy proposed in [14].

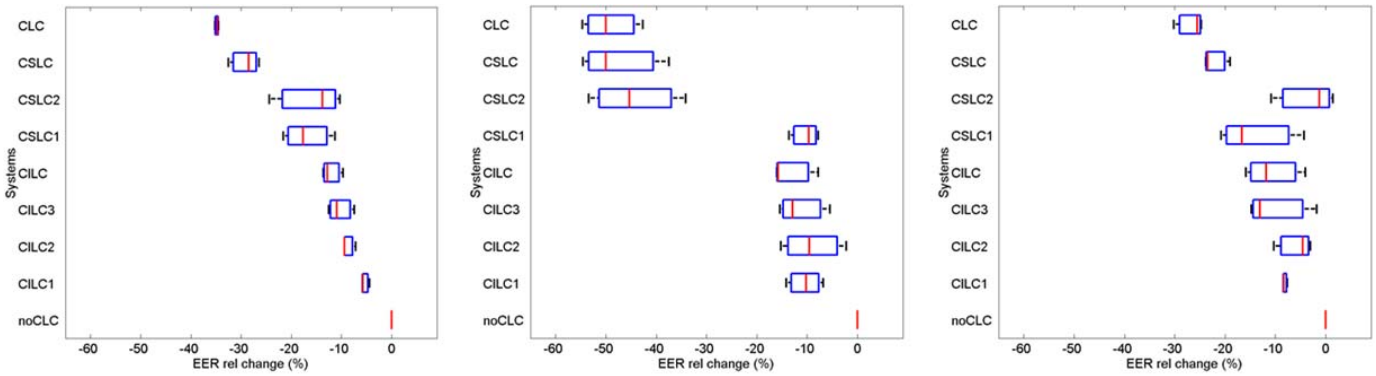
Now let us go back to Section IV-D1 about the potential of each individual modality on matching faces. By integrating CLC, the lowest EER (25.08%) is achieved by using the RGB modality with LBP on KinectFaceDB. While on the RGB-D-T, the best performance (14.96%) is obtained by using the thermal modality with HOG. Observe that with thermal, using LBP achieves much higher EER (23.14%) than HOG. This again demonstrates that HOG is more effective than LBP for capturing temperature variations. As shown in Table IV, using the depth modality leads to inferior performance than RGB and T modalities on the RGB-D-T database. However, by using CLC, we achieve the best performance (16.07%) with depth maps when LBP is used as the facial feature.

TABLE VIII
MEAN EERS (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY USING DIFFERENT SYSTEMS ON KINECTFACEDB

Feature	LBP		HOG	
Modality	RGB	D	RGB	D
noCLC	**43.05**	**47.62**	**44.90**	**46.11**
CILC1	36.22	43.11	38.74	41.88
CILC2	36.69	43.19	38.74	41.03
CILC3	33.92	41.30	35.31	39.31
CILC	**33.74**	**41.03**	**35.21**	**39.47**
CSLC1	27.75	38.70	31.44	37.70
CSLC2	28.25	37.50	31.50	37.90
CSLC	**27.14**	**35.68**	**29.10**	**34.43**
CLC	**25.08**	**34.93**	**27.71**	**33.71**

Table VIII and Table IX show the absolute improvement on matching performance induced by using cohort. To better evaluate the impact of CLC, we compute the relative improvement of a system using cohort with respect to the performance of the baseline system without cohort. The evaluation measure is the relative change of EER used in [14], which is computed as:

$$\text{rel. change of EER} = \frac{\text{EER}_{cohort} - \text{EER}_{noCohort}}{\text{EER}_{noCohort}} \quad (7)$$



(a) RGB (b) Depth (c) Thermal

Fig. 10. Boxplot of the relative change of EER using different systems for the three modalities on the RGB-D-T database, using LBP.

TABLE IX
MEAN EERS (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY USING DIFFERENT SYSTEMS ON RGB-D-T

Feature	LBP			HOG		
Modality	RGB	D	T	RGB	D	T
noCLC	**26.72**	**31.73**	**31.57**	**30.80**	**32.26**	**30.79**
CILC1	25.30	28.37	28.99	26.68	29.75	26.32
CILC2	24.40	28.75	29.63	26.63	29.82	28.02
CILC3	23.98	28.12	28.42	25.78	29.09	26.51
CILC	**23.51**	**27.49**	**28.20**	**25.19**	**28.28**	**25.82**
CSLC1	22.24	28.48	27.28	25.65	30.90	27.62
CSLC2	22.34	17.54	30.36	18.53	23.86	16.91
CSLC	**18.90**	**16.61**	**24.64**	**16.81**	**22.12**	**15.90**
CLC	**17.42**	**16.07**	**23.14**	**16.22**	**21.13**	**14.96**

TABLE X
MEAN EERS (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY USING RGB-D ON KINECTFACEDB

	noCLC	CILC	CSLC	CLC
LBP	38.00	31.48	26.02	**24.31**
HOG	41.07	34.21	29.41	**27.55**

where EER_{cohort} is the EER of a system using cohort, while $EER_{noCohort}$ is the EER of the baseline system. A negative change in the EER implies an improvement over the baseline system. Since there are three experiments corresponding to three folds, we summarize the results in a boxplot. Fig. 10 reports the results for the three modalities on RGB-D-T using the LBP feature. From these figures, we can clearly observe the relative contribution of each cohort coefficient to our pair matching task.

5) Fusion of Different Modalities: Finally, we fuse different modalities to see the improvement achieved by our multimodal complementary cohort strategy. Here, $m = 2$ for the KinectFaceDB database, while the value of m is 3 for the RGB-D-T database. The results on the two databases are listed in Table X and Table XI, respectively. By comparing Table X to Table VIII, we find in most cases, using RGB-D leads to better performance than using either of them on KinectFaceDB. For both LBP and HOG, the best results are obtained by our multimodal cohort strategy, i.e., RGB-D with CLC. Their corresponding EERs are 24.31% and 27.55%. Similarly, by comparing Table XI to Table IX, using RGB-D-T with CLC achieves the lowest EERs, which are 12.31% and 14.16% for LBP and HOG, respectively.

V. ANALYSIS OF COHORT LIST COMPARISON

In this section, we discuss several issues involved in our cohort investigation algorithm including its differences from existing cohort algorithms and the complementarity of CILC to CSLC for our face pair matching problem. To the best of our knowledge, this work, for the first time, introduces cohort investigation to multimodal face pair matching, as all the existing algorithms using cohort focus on the widely used RGB or intensity modality. Furthermore, by incorporating CILC (extracting cohort coefficients via sorted cohort positions) and CSLC (extracting cohort information via sorted cohort scores), we can comprehensively exploit the fixed cohort set. For performing CSLC, we employ exactly the picture-specific cohort ordering strategy proposed in [14]. When doing cohort identity list comparison, we borrow a similar idea to the Doppelgänger list comparison developed in [15]. Next, we list several differences between our CILC and the Doppelgänger list comparison.

As mentioned in Section III-A, in the Doppelgänger list, for each cohort identity, only one cohort face sample (the closest one to the test face image) is considered to calculate the similarity. This similarity is actually equal to our $sim1$ in $coefCI = [sim1, sim2, sim3]$. However, there are a set of face images for each cohort identity in our cohort set. Thus, each cohort identity can appear multiple times corresponding to multiple positions in the ranked cohort list. To compute similarity between such cohort lists, we designed two additional algorithms resulting in another two similarity measures $sim2$ and $sim3$. Another difference lies in the similarity normalization. As noted in Alg. 1, for computing all $sim1$, $sim2$ and $sim3$, we employ a divisor, i.e., $K \times K$. By doing this, we actually perform a normalization on the similarity. A further normalization procedure is followed by dividing the number of accumulated contributions. There is no such normalization operations in [15]. Besides, in their approach, only the extracted cohort information was employed

TABLE XI

MEAN EERS (%) OF 3-FOLD CROSS VALIDATION EXPERIMENTS BY FUSING DIFFERENT MODALITIES ON RGB-D-T

Feature	LBP				HOG			
Modality	RGB-D	RGB-T	D-T	RGB-D-T	RGB-D	RGB-T	D-T	RGB-D-T
noCLC	27.01	26.28	30.37	26.35	31.07	30.85	30.81	30.85
CILC	22.66	22.11	25.04	21.56	24.48	23.67	24.60	23.31
CSLC	13.05	19.18	14.28	13.12	16.74	15.12	15.90	15.18
CLC	12.39	16.95	13.46	**12.31**	15.91	14.18	14.74	**14.16**

TABLE XII

CLASSIFIER OUTPUTS AND THRESHOLDS FOR THE GENUINE AND IMPOSTOR PAIRS IN FIG. 5

	Probability			Threshold		
System	noCLC	CILC	CSLC	noCLC	CILC	CSLC
gPair	0.1964	0.9230	0.7576	0.5659	0.8096	0.8593
iPair	0.3281	0.3593	0.0395			

for classification without the raw matching score taken into account. We in fact incorporate $coefCI$ into $rawSC$ to assist the matching. This is consistent with the notion that the raw matching score contains much more information to drive the matching than cohort coefficients. Finally, the Doppelgänger list is computed using a large Library including 750,000 images from 337 subjects. While in our approach, there are only 306/324 images from 17/18 cohort subjects on the KinectFaceDB database and 1,700 images from 17 cohort subjects on the RGB-D-T database.

As reported in our experiments, the amount of reduced EER achieved by CILC is much less than that brought about by CSLC. By integrating CILC to CSLC, we observe that cohort coefficients extracted by CILC indeed contain some complementary information to $coefCS$. We take an example for this. In Table XII, we list the output of our logistic regression classifier on the RGB-D-T database, i.e., the probability of being a genuine pair for the two test pairs in Fig. 5. Together with the classifier output, we list also its threshold. Manifestly, a larger probability than the corresponding threshold leads to a genuine pair, whereas a lower one corresponds to an impostor pair. We use "gPair" (genuine pair) to denote$\{faceA1, faceA2\}$ and "iPair" (impostor pair) to represent $\{faceA1, faceB1\}$. LBP is used as the facial feature. For the genuine pair, using the direct matching score gets a probability of 0.1964, while the threshold is 0.5659. Obviously, the classifier will wrongly classify it into an impostor pair. By CILC, both the probability and the threshold increase. However, the probability increases much more than the threshold, leading to a higher probability than the threshold and thus a right decision. Similarly, the probability and the threshold increase after using CSLC, whereas the probability does not increase enough to exceed the increased threshold, thus resulting in still a wrong decision. For the impostor pair, all the three systems can achieve a right decision.

VI. CONCLUSION AND FUTURE WORK

In this paper, to handle large facial variation, we addressed the face pair matching issue by fusing different face modalities. By doing this, we can reduce the impact of diverse degrading factors, which usually affect different modalities in different degrees, on face matching performance. For the lack of representative information due to the few available face images, we proposed to further exploit a cohort set for additional information to better drive the matching. On two recently organized multimodal face databases, we investigated the power of each individual modality on matching faces and the performance achieved by fusing different modalities. Further, a set of experiments were performed to discover how much useful information the developed cohort investigation scheme can provide for the final matching. We observed that with different individual modalities, we got different face pair matching performance. By applying our multimodal complementary cohort strategy, we obtained promising results on both databases.

As facial biometric systems are expected to operate under challenging conditions, fusing different modalities and employing cohort information certainly offer two promising alternatives to render them more robust. For taking full advantage of different modalities, there is a great demand for developing modality-specific facial representations. It is equally important to design facial representations which take into account the correlation among different modalities. In addition, cohort coefficients discovered by CILC provide limited information for driving the matching, as seen from our experimental results. To better exploit cohort information from sorted cohort identities, further research might benefit from developing more effective similarity measures between sorted cohort lists and employing a much larger cohort set including large number of cohort identities.

ACKNOWLEDGMENT

The authors would like to thank Dr. Neslihan Kose from the Department of Multimedia Communications, EURECOM, for providing us with the KinectFaceDB database. Sincere thanks would be also given to the associate editor and anonymous reviewers for their 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, 2003.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[3] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[4] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. New York, NY, USA: Springer, 2008.

[5] G. B. Huang, M. Mattar, T. Berg, and E. L. Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2008, vol. 1, no. 2.

[6] 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.

[7] 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.

[8] Z. Lei, M. Pietikäinen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[10] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[11] 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.

[12] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. 10th Asian Conf. Comput. Vis.*, 2011, pp. 709–720.

[13] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[14] M. Tistarelli, Y. Sun, and N. Poh, "On the use of discriminative cohort score normalization for unconstrained face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2063–2075, Dec. 2014.

[15] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie, "Pose, illumination and expression invariant pairwise face-similarity measure via Doppelgänger list comparison," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2494–2501.

[16] Z. Chai, Z. Sun, H. Mendez-Vazquez, R. He, and T. Tan, "Gabor ordinal measures for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 14–26, Jan. 2014.

[17] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Comput. Vis. Image Understand.*, vol. 101, no. 1, pp. 1–15, 2006.

[18] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face recognition by fusing thermal infrared and visible imagery," *Image Vis. Comput.*, vol. 24, no. 7, pp. 727–742, 2006.

[19] 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, 1992.

[20] 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.

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

[22] N. Poh, A. Merati, and J. Kittler, "Making better biometric decisions with quality and cohort information: A case study in fingerprint verification," in *Proc. IEEE 17th Eur. Conf. Signal Process.*, Aug. 2009, pp. 70–74.

[23] J. Kittler, N. Poh, and A. Merati, "Cohort based approach to multiexpert class verification," in *Proc. 10th Int. Workshop Multiple Classifier Syst.*, vol. 6713. 2011, pp. 319–329.

[24] 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. 4, pp. 1270–1277, Aug. 2012.

[25] Y. Sun, C. B. Fookes, N. Poh, and M. Tistarelli, "Cohort normalization based sparse representation for undersampled face recognition," in *Proc. 11th Asian Conf. Comput. Vis. Workshops*, 2012, pp. 1–14.

[26] Y. Sun, M. Tistarelli, and N. Poh, "Picture-specific cohort score normalization for face pair matching," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep./Oct. 2013, pp. 1–8.

[27] 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.

[28] 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.

[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[30] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.

[31] M. Yang, L. Van Gool, and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 689–696.

[32] 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.

[33] 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.

[34] 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.

[35] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–12.

[36] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *Int. J. Robot. Res.*, vol. 31, no. 5, pp. 647–663, 2012.

[37] D. Holz, S. Holzer, R. Rusu, and S. Behnke, "Real-time plane segmentation using RGB-D cameras," in *Proc. Robot Soccer World Cup XV*, 2012, pp. 306–317.

[38] A. Ramey, V. González-Pacheco, and M. A. Salichs, "Integration of a low-cost RGB-D sensor in a social robot for gesture recognition," in *Proc. Int. Conf. Human-Robot Interact.*, 2011, pp. 229–230.

[39] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.

[40] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 186–192.

[41] M. Pamplona Segundo, S. Sarkar, D. Goldgof, L. Silva, and O. Bellon, "Continuous 3D face authentication using RGB-D cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 64–69.

[42] G. Goswami, M. Vatsa, and R. Singh, "RGB-D face recognition with texture and attribute features," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014.

[43] O. Nikisins, K. Nasrollahi, M. Greitans, and T. B. Moeslund, "RGB-D-T based face recognition," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1716–1721.

[44] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognit. Lett.*, vol. 24, no. 13, pp. 2115–2125, 2003.

[45] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," *IEEE Trans. Comput.*, vol. C-22, no. 11, pp. 1025–1034, Nov. 1973.

[46] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 841–848.

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

[48] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[49] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2007.

[50] N. Poh *et al.*, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 849–866, Dec. 2009.

[51] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, 2nd ed. London, U.K.: Springer-Verlag, 2009.

[52] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer Vision Using Local Binary Patterns* (Computational Imaging and Vision), vol. 40. London, U.K.: Springer, 2011.

[53] L. Zhang, B. Wu, and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[54] J. Choi, S. Hu, S. S. Young, and L. S. Davis, "Thermal to visible face recognition," in *Proc. SPIE, Sens. Technol. Global Health, Military Med., Disaster Response, Environ. Monitoring II Biometric Technol. Human Identification IX*, vol. 8371, p. 83711L, May 2012.

Yunlian Sun received the M.E. degree in computer science and technology from the Harbin Institute of Technology, China, in 2010, and the Ph.D. degree in *ingegneria elettronica, informatica e delle telecomunicazioni* from the University of Bologna, Italy, in 2014. She is currently a Postdoctoral Researcher with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. Her research interests focus on biometrics, pattern recognition, and computer vision.

Kamal Nasrollahi received the M.Sc. degree in computer engineering from Tehran Polytechnic, in 2007, and the Ph.D. degree in electrical engineering from Aalborg University, in 2010, with a focus on computer vision. He is currently an Associate Professor with the Visual Analysis of People Laboratory, Aalborg University, Denmark. He has been involved in five international research projects. His research interests include facial analysis systems, biometrics recognition, soft biometrics, and inverse problems. He has received an IEEE Conference Best Paper Award.

Zhenan Sun received the B.E. degree in industrial automation from the Dalian University of Technology, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), China, in 2006. Since 2006, he has been with NLPR, CASIA, as a Faculty Member. He is currently a Professor with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, CASIA. He is a member of the IEEE Computer Society, and the IEEE Signal Processing Society. His current research interests include biometrics, pattern recognition, and computer vision. He has authored or coauthored over 100 technical papers. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE BIOMETRICS COMPENDIUM.

Tieniu Tan (F'04) received the B.Sc. degree from Xian Jiaotong University, China, in 1984, and the M.Sc. and Ph.D. degrees from the Imperial College of Science, Technology and Medicine, London, U.K., in 1986 and 1989, respectively, all in electronics engineering. He is currently a Professor with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. His current research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a fellow of the International Association of Pattern Recognition.