Heterogeneous Face Recognition Using Kernel Prototype Similarities

Brendan F. Klare, *Member*, *IEEE*, and Anil K. Jain, *Fellow*, *IEEE*

Abstract—Heterogeneous face recognition (HFR) involves matching two face images from alternate imaging modalities, such as an infrared image to a photograph or a sketch to a photograph. Accurate HFR systems are of great value in various applications (e.g., forensics and surveillance), where the gallery databases are populated with photographs (e.g., mug shot or passport photographs) but the probe images are often limited to some alternate modality. A generic HFR framework is proposed in which both probe and gallery images are represented in terms of nonlinear similarities to a collection of prototype face images. The prototype subjects (i.e., the training set) have an image in each modality (probe and gallery), and the similarity of an image is measured against the prototype images from the corresponding modality. The accuracy of this nonlinear prototype representation is improved by projecting the features into a linear discriminant subspace. Random sampling is introduced into the HFR framework to better handle challenges arising from the small sample size problem. The merits of the proposed approach, called prototype random subspace (P-RS), are demonstrated on four different heterogeneous scenarios: 1) near infrared (NIR) to photograph, 2) thermal to photograph, 3) viewed sketch to photograph, and 4) forensic sketch to photograph.

Index Terms—Heterogeneous face recognition, prototypes, nonlinear similarity, discriminant analysis, local descriptors, random subspaces, thermal image, infrared image, forensic sketch

1 Introduction

An emerging topic in face recognition is matching between heterogeneous image modalities. Coined *heterogeneous face recognition (HFR)* [1], the scenario offers potential solutions to many difficult face recognition scenarios. While heterogeneous face recognition can involve matching between any two imaging modalities, the majority of scenarios involve a gallery dataset consisting of visible light photographs. Probe images can be of any other modality, though the practical scenarios of interest to us are infrared images (NIR and thermal) and hand-drawn facial sketches.

The motivation behind heterogeneous face recognition is that circumstances exist in which only a particular modality of a face image is available for querying a large database of mug shots (visible band face images). For example, when a subject's face can only be acquired in nighttime environments, the use of infrared imaging may be the only modality for acquiring a useful face image of the subject. Another example is situations in which no imaging system was available to capture the face image of a suspect during a criminal act. In this case a forensic sketch, drawn by a police artist based on a verbal description provided by a witness or the victim, is likely to be the only available source of a face image. Despite significant progress in the accuracy of face recognition systems [2], most commercial off-the-shelf (COTS) face recognition systems (FRS) are not designed to handle HFR scenarios. The need for face recognition systems specifically designed for the task of matching heterogeneous face images is of substantial interest.

This paper proposes a unified approach to heterogeneous face recognition that

1. achieves leading accuracy on multiple HFR scenarios,
2. does not necessitate feature descriptors that are invariant to changes in image modality,
3. facilitates recognition using different feature descriptors in the probe and gallery modalities, and
4. naturally extends to additional HFR scenarios due to properties 2 and 3 above.

2 Related Work

2.1 Heterogeneous Face Recognition

A flurry of research has emerged providing solutions to various heterogeneous face recognition problems. This began with sketch recognition using viewed sketches,[1] and has continued into other modalities such as near-infrared (NIR) and forensic sketches. In this section, we will highlight a representative selection of studies in heterogeneous face recognition as well as studies that use kernel-based approaches for classification.

Tang et al. spearheaded the work in heterogeneous face recognition with several approaches to synthesize a sketch from a photograph (or vice versa) [3], [4], [5]. Tang and Wang initially proposed an eigen-transformation method

1. A viewed sketch is a facial sketch drawn while viewing a photograph of the subject. The scenario is not practical because the photograph itself could be queried in the FR system.

- *B.F. Klare is with Noblis, 3150 Fairview Park Drive, Falls Church, VA 22042. E-mail: brendan.klare@noblis.org.*
- *A.K. Jain is with the Department of Computer Science and Engineering, Michigan State University, Room 3115, 428 S. Shaw Lane, Engineering Building, East Lansing, MI 48824-1226. E-mail: jain@cse.msu.edu.*

Manuscript received 18 Dec. 2011; revised 23 July 2012; accepted 16 Sept. 2012; published online 12 Oct. 2012.
Recommended for acceptance by M. Tistarelli.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-12-0905.
Digital Object Identifier no. 10.1109/TPAMI.2012.229.

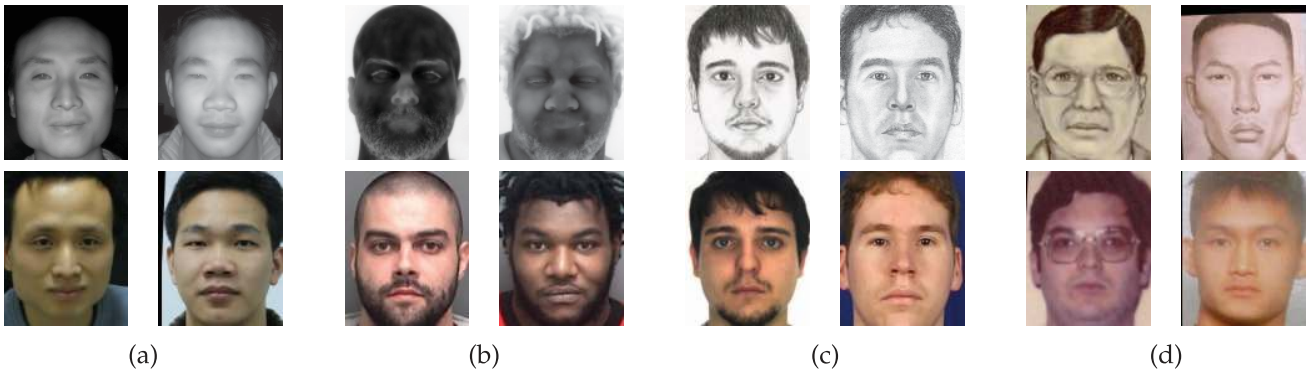


Fig. 1. Example images from each of the four heterogenous face recognition scenarios tested in our study. The top row contains probe images from (a) near-infrared, (b) thermal infrared, (c) viewed sketch, and (d) forensic sketch modalities. The bottom row contains the corresponding gallery photograph (visible band face image, called VIS) of the same subject.

[3]. Later, Liu et al. performed the transformation using local linear embedding to estimate the corresponding photo patch from a sketch patch [4]. Wang and Tang proposed a Markov random field model for converting a sketch into a photograph [5]. Other synthesis methods have been proposed as well [6], [7]. The generative transformation-based approaches have generally been surpassed in performance by discriminative feature-based approaches. A key advantage of synthesis methods is that once a sketch has been converted to a photograph, matching can be performed using existing face recognition algorithms. The proposed prototype framework is similar in spirit to these methods in that no direct comparison between face images in the probe and gallery modalities is needed.

A number of discriminative feature-based approaches to HFR have been proposed [8], [9], [10], [11], [12] which have shown good matching accuracies in both the sketch and NIR domains. These approaches first represent face images using local feature descriptors, such as variants of local binary patterns (LBPs) [13] and SIFT descriptors [14]. Liao et al. first used this approach on NIR to VIS face recognition by processing face images with a difference of Gaussian (DoG) filter, and encoding them using multiblock local binary patterns (MB-LBPs). Gentle AdaBoost feature selection was used in conjunction with R-LDA to improve the recognition accuracy. Klare and Jain followed this work on NIR to VIS face recognition by also incorporating SIFT feature descriptors and an RS-LDA scheme [10]. Bhatt et al. introduced an extended uniform circular local binary pattern to the viewed sketch recognition scenario [11]. Klare et al. encoded both viewed sketches and forensic sketches using SIFT and MLBP feature descriptors, and performed local feature-based discriminant analysis (LFDA) to improve the recognition accuracy [9]. Yi et al. [15] offered a local patch-based method to perform HFR on partial NIR face images. Zhang et al. extracted local features and performed recognition between sketches and photos using coupled information-theoretic encoding [16]. Lei and Li applied coupled spectral regression (CSR) for NIR to VIS recognition [12]. In [12], CSR was extended to Kernel CSR, which is similar to the proposed prototype representation in this work.

The synthesis method by Li et al. is the only known method to perform recognition between thermal IR and visible face images [17]. The only method to perform recognition between forensic sketches and visible face images is Klare et al. [9], which is also one of two methods, to our knowledge, that has been tested on two different HFR scenarios (viewed sketch and forensic sketch). The other method is Lin and Tang's [18] common discriminant recognition framework, which was applied to viewed sketches and near-infrared images. In this work, the proposed prototype random subspace (P-RS) framework is tested on four different HFR scenarios.

2.2 Kernel Prototype Representation

The core of the proposed approach involves using a relational feature representation for face images (illustrated in Fig. 2). By using kernel similarities between a novel face pattern and a set of prototypes, we are able to exploit the kernel trick [19], which allows us to generate a high dimensional, nonlinear representation of a face image using compact feature vectors.

The benefit of a prototype-based approach is provided by Balcan et al. [19]. Given access to the data distribution and a kernel similarity function, a prototype representation

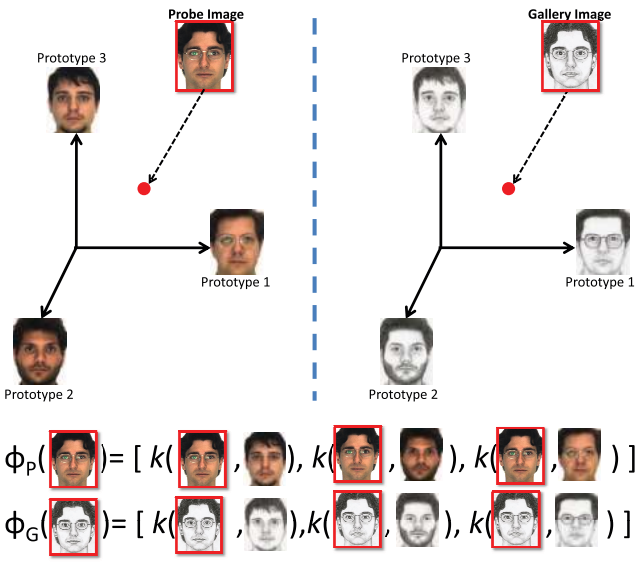


Fig. 2. The proposed face recognition method describes a face as a vector of kernel similarities to a set of prototypes. Each prototype has one face image in the probe and gallery modalities.

is shown to approximately maintain the desired properties of the high-dimensional kernel space in a more efficient representation by using the kernel trick. While it is not common to refer to kernel methods as prototype representations, in this work we emphasize the fact that kernel methods use a training set of images (which serve as prototypes) to implicitly estimate the distribution of the nonlinear feature space. One key to our framework is that each prototype has a pattern for each image modality.

The proposed kernel prototype approach is similar to the object recognition method of Quattoni et al. [20]. Kernel PCA [21] and Kernel LDA [22], [23] approaches to face recognition have used a similar approach, where a face is represented as the kernel similarity to a collection of prototype images in a high-dimensional space. The biometric indexing scheme by Gyaourova and Ross used similarity scores to a fixed set of references in the face and fingerprint modality [24].

These prior works differ from the proposed method because only a single prototype is used per training subject. By contrast, our approach is designed for heterogeneous face recognition, and uses two prototype images per subject (one per modality). Our earlier work [25] utilized a similar approach that did not exploit the benefit of nonlinear kernels, but did use a separate pattern from each image modality (sketch and photo) for each prototype. The kernel coupled spectral regression by Lei and Li used a similar approach of representing heterogeneous face images as nonlinear similarities to a set of prototypes [12].

2.3 Proposed Method

The proposed method presents a new approach to heterogeneous face recognition, and extends existing methods in face recognition. The use of a nonlinear similarity representation is well suited to the HFR problem because a set of training subjects with an image from each modality can be used as the prototypes and, depending on the modality of a new image (probe or gallery), the image from each prototype subject can be selected from the corresponding modality. Unlike previous feature-based methods, where an image descriptor invariant to changes between the two HFR modalities was needed, the proposed framework only needs descriptors that are effective within each domain. Further, the proposed method is effective even when different feature descriptors are used in the probe and gallery domains. The proposed prototype framework is described in detail in Section 4.

The accuracy of the HFR system is improved using a random subspace framework in conjunction with linear discriminant analysis (LDA), as described in Section 5. The previous (or baseline) method of feature-based random subspaces [10] is revisited in Section 6. Experimental results on four different heterogeneous face recognition scenarios (thermal, near-infrared, viewed sketch, and forensic sketch) are provided in Section 7, and all the results are benchmarked with a commercial face matcher.

While we demonstrate the strength of the proposed framework on many different HFR scenarios, the parameters controlling the framework are the same across all tested scenarios. This shows that the contribution of this work is a generic framework for improving solutions to the general HFR problem. Future use of the proposed framework will benefit from selecting parameters tailored to a specific scenario; however, that is beyond the scope of this work.

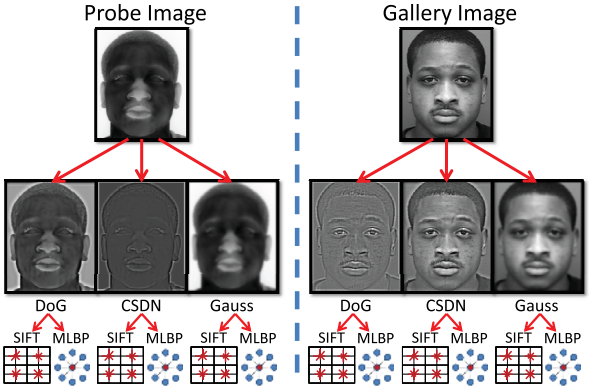


Fig. 3. Example of thermal probe and visible gallery images after being filtered by a difference of Gaussian, center surround divisive normalization, and Gaussian image filter. The SIFT and MLBP feature descriptors are extracted from the filtered images, and kernel similarities are computed within this image descriptor representation.

3 Image Preprocessing and Representation

All face images are initially represented using a feature-based representation. The use of local feature descriptors has been argued to closely resemble the postulated representation of the human visual processing system [26], and they have been shown to be well suited for face recognition [27].

3.1 Geometric Normalization

The first step in representing face images using feature descriptors is to geometrically normalize the face images with respect to the location of the eyes. This step reduces the effect of scale, rotation, and translation variations. The eye locations for the face images from all modalities are automatically estimated using Cognitec's FaceVACS SDK [28]. The only exceptions are the thermal face images where the eyes are manually located for both the proposed method and the FaceVACS baseline.

Face images are geometrically normalized by 1) performing planar rotation to set the angle between the eyes to 0 degrees, 2) scaling the images so that the distance between the two pupils is 75 pixels, and 3) cropping the images to a height of 250 pixels and a width of 200 pixels, with the eyes horizontally centered and vertically placed at row 115.

3.2 Image Filtering

Face images are filtered with three different image filters. These filters are intended to help compensate for both intensity variations within an image domain (such as nonuniform illumination changes), as well appearance variations between image domains. The second aspect is of particular importance for the direct random subspace (D-RS) framework (see Section 6). An example of the effects of each image filter can be seen in Fig. 3.

The three image filters used are as follows.

3.2.1 Difference of Gaussian

A difference of Gaussian image filter has been shown by Tan and Triggs to improve face recognition performance in

the presence of varying illumination [29], as well as in an NIR to VIS matching scenario by Liao et al. [8]. A difference of Gaussian image is generated by convolving an image with a filter obtained by subtracting a Gaussian filter of width σ_1 from a Gaussian filter of width σ_2 ($\sigma_2 > \sigma_1$). In this paper, $\sigma_1 = 2$ and $\sigma_2 = 4$.

3.2.2 Center-Surround Divisive Normalization (CSDN)

Meyers and Wolf [30] introduced the center-surround divisive normalization filter in conjunction with their biologically inspired face recognition framework. The CSDN filter divides the value of each pixel by the mean pixel value in the $s \times s$ neighborhood surrounding the pixel. The nonlinear nature of the CSDN filter is seen as a compliment to the DoG filter. In our implementation, $s = 16$.

3.2.3 Gaussian

The Gaussian smoothing filter has long been used in image processing applications to remove noise contained in high spatial frequencies while retaining the remainder of the signal. The width of the filter used in our implementation was $\sigma = 2$.

3.3 Local Descriptor Representation

Once an image is geometrically normalized and filtered using one of the three filters, local feature descriptors are extracted from uniformly distributed patches across the face. In this work, we use two different feature descriptors to represent the face image: the SIFT descriptor [14] and Local Binary Patterns [13]. The SIFT feature descriptor has been used effectively in face recognition [27], sketch to VIS matching [9], and NIR to VIS matching [10]. LBP features have a longer history of successful use in face recognition. Ahonen et al. originally proposed their use for face recognition [31], Li et al. demonstrated their use in NIR to NIR face matching [32], and they have also been successfully applied to several HFR scenarios [8], [9], [10], [11].

The SIFT and LBP feature representations are effective in describing face images due to their ability to encode the structure of the face and their stability in the presence of minor external variations [27]. Each feature descriptor describes an image patch as a d-dimensional vector that is normalized to sum to one. The face image is divided into a set of N overlapping patches of size 32×32. Each patch overlaps its vertical and horizontal neighbors by 16 pixels. With a face image of size 200×250, this results in a total of 154 total patches.

Multiscale local binary patterns (MLBP) [9], a variant of the LBP descriptor, is used in place of LBP in this work. MLBP is the concatenation of LBP feature descriptors with radii $r = \{1, 3, 5, 7\}$.

Let I be a (normalized and filtered) face image. Let $f_{F,D}(I, a)$ denote the local feature descriptor extracted from image I at patch a, $1 \leq a \leq N$, using image filter F and feature descriptor D. The DoG, CSDN, and Gaussian image filters are, respectively, referred to as F_d, F_c, and F_g. The MLBP and SIFT descriptors are, respectively, referred to as D_m and D_s. Using 16 histograms and 8 orientation bins, as described by Lowe [14], the SIFT descriptor yields a 128D feature descriptor. Using uniform patterns at eight sampling locations, as decribed by Ojala et al. [13], the LBP descriptor yields a 59D feature descriptor. This results in a 236D MLBP feature descriptor ($f_{F,D_m}(I, a) \in \mathbb{R}^{236}$). Finally, we have

$$f_{F,D}(I) = [f_{F,D}(I,1)^{\mathrm{T}},\ \ldots,\ f_{F,D}(I,N)^{\mathrm{T}}]^{\mathrm{T}}, \tag{1}$$

which is the concatenation of all N feature descriptors. Thus, $f_{F,D_s}(I) \in \mathbb{R}^{128 \cdot N}$ and $f_{F,D_m}(I) \in \mathbb{R}^{236 \cdot N}$.

Using the three filters and two descriptors, we have six different representations available for face image I, namely, $f_{F_d,D_m}(I)$, $f_{F_c,D_m}(I)$, $f_{F_g,D_m}(I)$, $f_{F_d,D_s}(I)$, $f_{F_c,D_s}(I)$, and $f_{F_g,D_s}(I)$.

4 Heterogeneous Prototype Framework

4.1 Prototype Representation

The heterogeneous prototype framework begins with images from the probe and gallery modalities represented by (possibly different) feature descriptors for each of the N image patches, as described in the previous section. For compactness, let $f(I)$ represent $f_{F,D}(I)$. The similarity between two images is measured using a kernel function $k : f(I)\ \mathrm{x}\ f(I) \rightarrow \mathbb{R}$.

Let $\mathcal{T}$ be a set of training images consisting of n_t subjects. The training set contains a probe image P_i and gallery image G_i for each of the n_t subjects. That is,

$$\mathcal{T} = \{P_1, G_1, \ldots, P_{n_t}, G_{n_t}\}, \tag{2}$$

For both the probe and gallery modalities, two positive semi-definite kernel matrices K^P and K^G are computed between the training subjects. The probe kernel matrix is $K^P \in \mathbb{R}^{n_t, n_t}$, and the gallery kernel matrix is $K^G \in \mathbb{R}^{n_t, n_t}$. The entries in the ith row and jth column of K^P and K^G are

$$K^P(i,j) = k(f(P_i) \quad f(P_j)), \tag{3}$$

$$K^G(i,j) = k(\ f(G_i) \quad f(G_j)), \tag{4}$$

where $k(\cdot,\cdot)$ is the kernel similarity function. Results in all experiments in this work use the cosine kernel function:

$$k(f(P_i), f(G_i)) = \frac{\langle f(P_i), f(G_i) \rangle}{\|f(P_i)\| \cdot \|f(G_i)\|}. \tag{5}$$

The cosine kernel was chosen because it resulted in consistently higher accuracy on all tested scenarios compared to the radial basis function kernel and the polynomial kernel. Additionally, we preferred the cosine kernel because it is devoid of parameters.

Let P and G, respectively, be test probe and gallery face images, i.e., $(P, G \notin \mathcal{T})$. The function $\phi_P(P)$ returns a vector containing the kernel similarity of image P to each image P_i in $\mathcal{T}$. For gallery image G, $\phi_G(G)$ returns a vector of kernel similarities to the gallery prototypes G_i. Thus, face images are represented as the relational vector $\phi_P(P) \in \mathbb{R}^{n_t}$ for a probe image and $\phi_G(G) \in \mathbb{R}^{n_t}$ for a gallery image. More precisely, we have

$$\phi_P(P) = [k(f(P), f(P_1)), \ldots, k(f(P), f(P_{n_t}))]^{\mathrm{T}}, \tag{6}$$

$$\phi_G(G) = [k(f(G), f(G_1)), \ldots, k(f(G), f(G_{n_t}))]^{\mathrm{T}}. \tag{7}$$

Using this prototype-based representation, extreme inputs to the system (e.g., a nonface image) will cause the

kernel similarity to degenerate to the kernel minimum (0 in the case of the cosine kernel). This allows the system to remain stable with respect to scale.

Additionally, because the feature vectors $\phi_P(P)$ and $\phi_G(G)$ are a measure of the similarity between the test image and the prototype training images, the feature spaces for similarity computation do not have to be the same for the probe and gallery modalities. For example, the probe images could be represented using $F_{F,D_s}(P)$ and the gallery images could be represented using $F_{F,D_m}(G)$. Despite the fact that the SIFT and MLBP feature descriptors are heterogeneous features, the relational representation allows them to be represented in a common feature space. This is based on the assumption that

$$k(f(P), f(P_i)) \approx k(f(G), f(G_i)). \tag{8}$$

We will next introduce a discriminant subspace technique to project these prototype features into a linear subspace that better satisfies (8). When necessary, the tersely presented notation of $\phi_P(I)$ or $\phi_G(I)$ will be expanded to the more verbose notation $\phi_P^{F,D}(I)$ or $\phi_G^{F,D}(I)$, respectively, in order to specify which feature descriptor and image filter is initially being used to represent the image I. For example, $\phi_P^{F_c,D_s}(I)$ denotes the prototype similarity of image I when represented using the CSDN image filter and SIFT descriptors.

4.2 Discriminant Analysis

After representing the images in the training set $\mathcal{T}$ in the aforementioned prototype representation, we next learn linear subspaces using linear discriminant analysis [33] to enhance the discriminative capabilities of the prototype representation $\phi(\cdot)$. LDA (and its variants) has consistently demonstrated its ability to improve the accuracy of various recognition algorithms through feature extraction and dimensionality reduction. The benefits of LDA in the context of face recognition have been demonstrated on image pixel representations [33], [34], Gabor features [35], and image descriptors [8], [9].

We learn the linear projection matrix W by following the conventional approach for high-dimensional data, namely, by first applying PCA, followed by LDA [33]. In all experiments, the PCA step was used to retain 99.0 percent of the variance. Let X be a matrix whose columns contain the prototype representation of each image in $\mathcal{T}$:

$$X = [\phi_P(P'_1), \phi_G(G'_1), \ldots, \phi_P(P'_{n_t}), \phi_G(G'_{n_t})]. \tag{9}$$

Let X' denote the mean-centered version of X. The initial step involves learning the subspace projection matrix W'_1 by performing principal component analysis (PCA) on X' to reduce the dimensionality of the feature space. Next, the within-class and between-class scatter matrices of ${W'_1}^{\mathrm{T}} \cdot X'$ (respectively), S_W and S_B, are computed. The dimension of the subspace W'_1 is such that S_W will be of full rank. The scatter matrices are built using each subject as a class; thus one image from the probe and gallery modality represents each class. A more detailed description of how to compute S_W and S_B is described in [9]. Last, the matrix W'_2 is learned by solving the generalized eigenvalue problem:

$$S_B \cdot W'_2 = \Lambda \cdot S_W \cdot W'_2. \tag{10}$$

This yields the LDA projection matrix W, where

$$W = \left({W'_2}^{\mathrm{T}} \cdot {W'_1}^{\mathrm{T}}\right)^{\mathrm{T}}. \tag{11}$$

Letting μ denote the mean of X, the final representation for an unseen probe or gallery image I using the prototype framework is $W^{\mathrm{T}} \cdot (\phi(I) - \mu)$. Subsequent uses of W in this work will assume the appropriate removal of the mean μ from $\phi(I)$ for terseness.

5 Random Subspaces

5.1 Motivation

The proposed heterogeneous prototype framework uses training data to define the prototypes and to learn the linear subspace projection matrix W. This requirement on training data raises two (somewhat exclusive) issues in the prototype representation framework. The first issue is that the number of subjects in $\mathcal{T}$ (i.e., the number of prototypes) is generally too small for an expressive prototype representation. While Balcan et al. demonstrated that the number of prototypes does not need to be large to approximately replicate the data distribution [19], their applications primarily dealt with binary classification and a small number of features. When applying a prototype representation to face recognition, a large number of classes (or subjects) and features are present. The small sample size problem implies that the number of prototypes needed to approximate the underlying data distribution should be large [36].

The second issue is also related to the small sample size problem [36]. This common problem in face recognition arises from too few training subjects to learn model parameters that are not susceptible to generalization errors. In the heterogeneous prototype framework this involves learning a W matrix that generalizes well.

A number of solutions exist to tackle the small sample size problem in face recognition. Most are designed to handle deficiencies in the subspace W, such as dual-space LDA [34] and direct LDA [37]. Regularization methods such as R-LDA [38] also address degenerative properties of W. However, these methods do not address the issue of too few prototypes for an expressive representation.

Another approach to handle deficiencies in learning parameters is the use of random subspaces [39]. The random subspace method samples a subset of features and performs training in this reduced feature space. Multiple sets (or bags) of randomly sampled features are generated, and for each bag the parameters are learned. This approach is similar to the classical bagging classification scheme [40], where the training instances are randomly sampled into bags multiple times and training occurs on each bag separately. Ensemble methods such as Ho's random subspaces [39] and Breiman's bagging classifiers have been demonstrated to increase the generalization of an arbitrary classifier [41].

Wang and Tang demonstrated the effectiveness of random sampling LDA (RS-LDA) for face recognition [42]. Their approach combined random subspaces and bagging by sampling both features and training instances. For each random sample space, a linear subspace was learned. Klare and Jain utilized this approach in the HFR

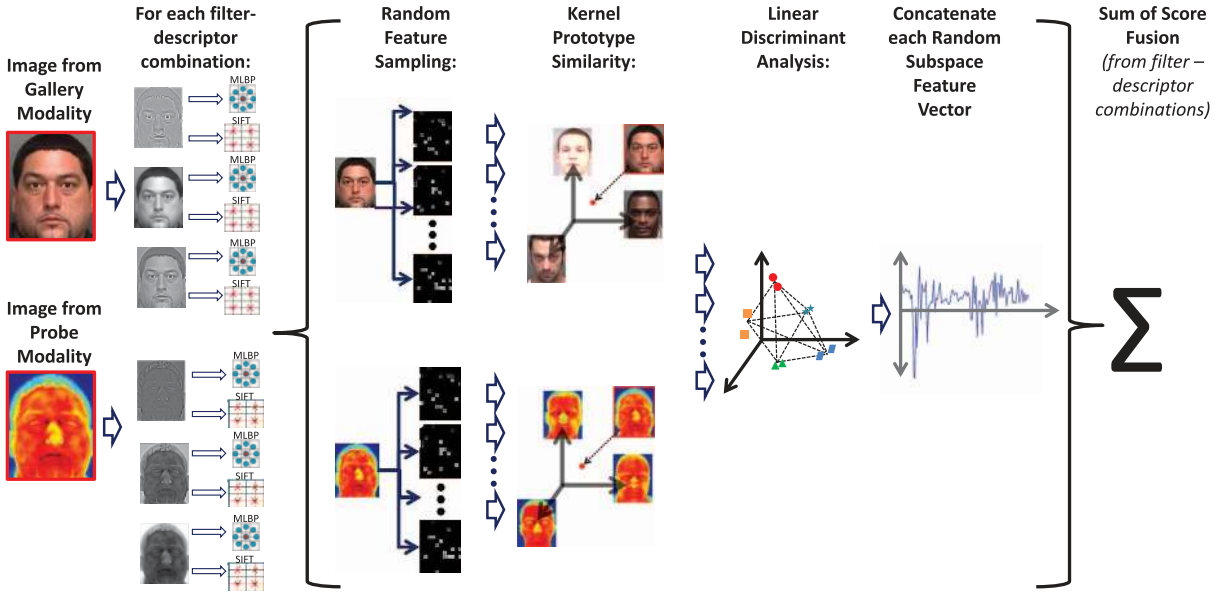


Fig. 4. Overview of the proposed algorithm for performing heterogenous face recognition using prototype similarities.

scenario of NIR/VIS by using multiple samplings of face patches described by local feature descriptors [10].

We consider random sampling ideal for the prototype recognition framework because it is able to satisfactorily address the two limitations: 1) The number of prototypes is multiplied by the number of bags, which improves the expressiveness of the prototype representation, and 2) the use of an ensemble method improves deficiencies in W. Further unification of these two separate problems into a single solution offers a simpler framework.

5.2 Prototype Random Subspaces

The prototype random subspace framework uses B different bags (or samples) of the N face patches. Each sample consists of $\alpha \cdot N$ patches, $0 \leq \alpha \leq 1$. For bag b ($b = 1 \dots B$), we have the integer vector $\kappa_b \in \mathbb{Z}^{\alpha \cdot N}$, where each component of κ_b is a unique randomly sampled value from $1 \dots N$. It is assumed that α is selected such that $\alpha \cdot N$ is an integer. An example of randomly sampled face patches in shown in Fig. 5.

Let $f(I, \kappa_b)$ denote the concatenation of the $\alpha \cdot N$ descriptors from the randomly selected patch indices in κ_b. That is,

$$f(I, \kappa_b) = [f(I, \kappa_b(1))^{\mathrm{T}}, \dots, f(I, \kappa_b(\alpha \cdot N))^{\mathrm{T}}]^{\mathrm{T}}. \quad (12)$$

Letting K_b^P and K_b^G denote the probe and gallery kernel similarity matrices for bag b, we modify (3) and (4) to

$$K_b^P(i,j) = k(f(P_i, \kappa_b) \quad f(P_j, \kappa_b)), \quad (13)$$

$$K_b^G(i,j) = k(f(G_i, \kappa_b) \quad f(G_j, \kappa_b)). \quad (14)$$

The preliminary prototype representation $\phi(\cdot)$ is now modified to $\phi(\cdot, \cdot)$ as

$$\phi_P(P, \kappa_b) = [\, k(f(P, \kappa_b), f(P_1, \kappa_b)), \dots, k(f(P, \kappa_b), f(P_{n_t}, \kappa_b))\,]^{\mathrm{T}}, \quad (15)$$

$$\phi_G(G, \kappa_b) = [\, k(f(G, \kappa_b), f(G_1, \kappa_b)), \dots, k(f(G, \kappa_b), f(G_{n_t}, \kappa_b))]^{\mathrm{T}}. \quad (16)$$

Linear discriminant analysis is performed separately for each bag. Using training set $\mathcal{T}$ (which was also used to define the prototypes), we learn B subspace projection matrices W_b, $b = 1 \dots B$.

A new face image I is represented in the random subspace prototype framework as $\Phi(I)$, where $\Phi(I)$ is the concatenation of each linearly projected prototype representation from each of the B random subspace bags. That is,

$$\Phi_P(P) = [(W_1^{\mathrm{T}} \cdot \phi_P(P, \kappa_1))^{\mathrm{T}}, \dots, (W_B^{\mathrm{T}} \cdot \phi_P(P, \kappa_B))^{\mathrm{T}}]^{\mathrm{T}}, \quad (17)$$

$$\Phi_G(G) = [(W_1^{\mathrm{T}} \cdot \phi_G(G, \kappa_1))^{\mathrm{T}}, \dots, \quad (W_B^{\mathrm{T}} \cdot \phi_G(G, \kappa_B))^{\mathrm{T}}]^{\mathrm{T}}. \quad (18)$$

For terseness we have omitted the superscript F and D in the previous equations. For example, in (17), $\Phi_P^{F,D}(I)$ is abbreviated to $\Phi(P)$ for image filter F and descriptor D to represent I.

A visual summary of the proposed prototype random subspaces method can be found in Fig. 4.

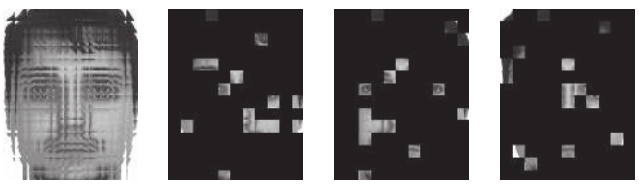


Fig. 5. The process of randomly sampling image patches is illustrated. (a) All image patches. (b), (c), and (d) Bags of randomly sampled patches. The kernel similarity between SIFT and MLBP descriptors at each patch of an incoming image and the prototypes of corresponding modality are computed for each bag. Images from [43].

5.3 Recognition

Given a probe face image P and a gallery face image G, we define their similarity $S(P,G)$ using the cosine similarity measure:

$$S(P,G) = \frac{\langle \Phi_P(P), \Phi_G(G) \rangle}{\|\Phi_P(P)\| \cdot \|\Phi_G(G)\|}. \tag{19}$$

Further, we let $S^{F_1,D_1}_{F_2,D_2}(P,G)$ denote the similarity between the probe P represented using filter F_1 and descriptor D_1, and gallery image G represented in terms of filter F_2 and descriptor D_2. That is,

$$S^{F_1,D_1}_{F_2,D_2}(P,G) = \frac{\langle \Phi^{F_1,D_1}_P(P),\ \Phi^{F_2,D_2}_G(G) \rangle}{\|\Phi^{F_1,D_1}_P(P)\| \cdot \|\Phi^{F_2,D_2}_G(G)\|}. \tag{20}$$

This similarity measure facilitates recognition using a threshold for a binary verification scenario or a nearest neighbor matcher for an identification scenario.

5.4 Score Level Fusion

The proposed framework naturally lends to fusion of the different feature representations. For example, given one image filter F and two feature descriptors D_1 and D_2, we can utilize the following sum of similarity scores between probe image P and gallery image G : $\{S^{F,D_1}_{F,D_1}(P,G) + S^{F,D_2}_{F,D_2}(P,G) + S^{F,D_2}_{F,D_1}(P,G) + S^{F,D_2}_{F,D_1}(P,G)\}$. Min-max score normalization is performed prior to fusion.

6 Baselines

6.1 Commercial Matcher

The accuracy of the proposed prototype random subspace framework is compared against Cognitec's FaceVACS [28] COTS FRS. Comparing the accuracy of our system against a leading COTS FRS offers an unbiased baseline of a state-of-the-art commercial matcher on each HFR scenario. FaceVACS was chosen because it is considered as one of the best commercial face matchers and, in our internal tests, it excels at HFR scenarios (with respect to other commercial matchers). For example, the accuracy of FaceVACS on NIR to VIS [10] and Viewed Sketch to VIS [9] performed on par with some previously published HFR methods.

6.2 Direct Random Subspaces

In addition to a commercial face recognition system, the proposed prototype recognition system is also compared against a recognition system that directly measures the difference between probe and gallery images using a common feature descriptor representation. As discussed previously, most recent approaches to heterogeneous face recognition involve directly measuring the similarity between two face images from alternate modalities using feature descriptors [8], [9], [10], [11].

The random subspace framework from [10] is used as the baseline because it is the most similar to the proposed prototype framework, thus helping to isolate the difference between using kernel prototype similarities versus directly measuring the similarity. Further, because most of the datasets tested in Section 7 are in the public domain, the proposed framework may also be compared against any other published method on these datasets.

To briefly summarize the direct random subspace approach using our notation, at each bag b the D-RS framework represents an image as $f_{F,D}(I,\kappa_b)$. LDA is performed on each bag to learn the projection matrix $\tilde{W}_b$ using the training set $\mathcal{T}$. The final representation $\Psi(\cdot)$ is the concatenation of the projected vector at each bag:

$$\Psi_{F,D}(I) = [(\tilde{W}_1^{\mathrm{T}} \cdot f_{F,D}(I,\kappa_1))^{\mathrm{T}}, \ldots, (\tilde{W}_B^{\mathrm{T}} \cdot f_{F,D}(I,\kappa_B))^{\mathrm{T}}]^{\mathrm{T}}. \tag{21}$$

The similarity between probe image P and gallery image G (each represented with filter F and descriptor D) is computed using the cosine similarity metric:

$$\tilde{S}_{F,D}(P,G) = \frac{\langle \Psi_{F,D}(P), \Psi_{F,D}(G) \rangle}{\|\Psi_{F,D}(P)\| \cdot \|\Psi_{F,D}(G)\|}. \tag{22}$$

Unlike P-RS, D-RS must use the same D for the probe and gallery images. This follows from the fact that $f_{f,D_1}(I)$ and $f_{f,D_2}(I)$ are of generally different dimensionality and also have a different interpretation.

The D-RS algorithm is presented as a baseline approach motivated by our earlier works [9], [10]. However, through directly measuring facial similarities, D-RS is able to capture additional information that may be useful in the recognition process. As such, it will be shown in Section 7 that score level fusion between D-RS and P-RS often provide the highest accuracies in most HFR scenarios.

D-RS will be used in conjunction with the six filter/descriptor representations presented in Section 3 (SIFT+DoG, MLBP+CSDN, etc.). Results will be presented from the sum-score fusion of the min-max normalized scores from these six representations.

7 Experiments

The results provided are based on the following parameter values: $\alpha = 0.1$ and $B = 30$. A cosine kernel was used to compute the prototype similarity and 99.0 percent of the variance was retained in the PCA step of LDA.

7.1 Databases

Five different matching scenarios are tested in this paper: four heterogeneous face recognition scenarios and one standard face recognition scenario. Example images from each of HFR dataset can be found in Fig. 1. Results shown on each dataset are the average and standard deviation of five random splits of training and testing subjects. In every experiment, no subject that was used in training was used for testing.

7.1.1 Dataset 1—Near-Infrared to Visible (Fig. 1a)

The first dataset consists of 200 subjects with probe images captured in the near-infrared spectrum (~780-1,100 nm) and gallery images captured in the visible spectrum. Portions of this dataset are publicly available for download.[2] This dataset was originally used by Li et al. [8], [32]. Only one NIR and one VIS image per subject are used, making the scenario more difficult than previous experiments which benefited from multiple images per subject in training and

2. http://www.cbsr.ia.ac.cn/english/Databases.asp.

gallery enrollment [8], [10]. The data was split as follows: $n_t = 133$ subjects were used for training set $\mathcal{T}$ and the remaining 67 subjects we used for testing.

7.1.2 Dataset 2—Thermal to Visible (Fig. 1b)

The second dataset is a private dataset collected by the Pinellas County Sheriff's Office (PCSO) and consists of 1,000 subjects with thermal infrared probe images and visible (mug shot) gallery images. The thermal infrared images were collected using a FLIR Recon III ObservIR camera, which has sensitivity in the range of 3-5 μm. The data was split as follows: $n_t = 667$ subjects were used for training set $\mathcal{T}$ and the remaining 333 subjects were used for testing.

7.1.3 Dataset 3—Viewed Sketch to Visible (Fig. 1c)

The third dataset is the CUHK sketch dataset,[3] which was used by Tang and Wang [3], [5]. The CUHK dataset consists of 606 subjects with a viewed sketch image for probe and a visible photograph for gallery. A viewed sketch is a hand-drawn sketch of a face which is drawn while looking at a photograph of the subject. The photographs in the CUHK dataset are from the AR [43], XM2VTS [44], and CUHK student [3], [5] datasets. The 606 subjects were split to form a training set $\mathcal{T}$ with $n_t = 404$ subjects, and the remaining 202 subjects were used for testing.

7.1.4 Dataset 4—Forensic Sketch to Visible (Fig. 1d)

The fourth and final heterogeneous face dataset consists of real-world forensic sketches and mug shot photos of 159 subjects. This dataset is described in [9]. Forensic sketches are drawn by an artist based only on an eye witness description of the subject. The forensic sketch dataset is a collection of images from Gibson [45], Taylor [46], the Michigan State Police, and the Pinellas County Sheriff's Office. Each sketch contains a suspect involved in a real crime and the mug shot photo was only available after the subject had later been identified. Forensic sketches contain incomplete information regarding the subject and are one of the most difficult HFR scenarios because the sketches often do not closely resemble the photograph. The number of subjects subjects used in $\mathcal{T}$ is 106, and 53 subjects are used for the test set.

7.1.5 Dataset 5: Standard Face Recognition

A fifth nonheterogeneous (i.e., homogeneous) dataset is used to demonstrate the ability of the proposed approach to operate in standard face recognition scenarios as well. The dataset consists of one probe and one gallery photograph of 876 subjects, where 117 subjects were from the AR dataset [43], 294 subjects were from the XM2VTS dataset [44], 193 subjects from the FERET dataset [47], and 272 subjects were from a private dataset collected at the University of Notre Dame. This is the same dataset used in [27].

7.1.6 Enlarged Gallery

A collection of 10,000 mug shot images from 10,000 different subjects was used in certain experiments to increase the size of the gallery. These mug shot images were provided by the Pinellas County Sheriff's Office. Any experiment using these additional images will have a gallery with the number of testing subjects plus 10,000 images. Experiments with a large gallery are meant to present results that more closely resemble real-world face retrieval scenarios that would occur in forensic and intelligence applications of heterogeneous face recognition.

3. The CUHK dataset is publicly available for download at http://mmlab.ie.cuhk.edu.hk/facesketch.html.

7.2 Results

Fig. 6 lists the rank retrieval results of P-RS, D-RS, and FaceVACS for each dataset using the additional 10,000 gallery images for each experiment. Additionally, the true accept rates (TAR) at false accept rates (FAR) of 1.0 and 0.1 percent are listed to facilitate comparisons with other published methods.

The results for P-RS are the fusion of the match scores from $\{S^{F_d,D_s}_{F_d,D_s} + S^{F_c,D_s}_{F_c,D_s} + S^{F_G,D_s}_{F_G,D_s} + S^{F_d,D_m}_{F_d,D_m} + S^{F_c,D_m}_{F_c,D_m} + S^{F_G,D_m}_{F_G,D_m}\}$, i.e., the same features are used in the probe and gallery images. Similarly, D-RS is the fusion of the match scores from $\{\tilde{S}_{F_d,D_s} + \tilde{S}_{F_c,D_s} + \tilde{S}_{F_G,D_s} + \tilde{S}_{F_d,D_m} + \tilde{S}_{F_c,D_m} + \tilde{S}_{F_G,D_m}\}$. Results from these same matchers are also displayed in CMC (cumulative match characteristic) plots in Fig. 7.

The CMC results of matching NIR face images to standard face images are shown in Fig. 7a. The accuracies in Fig. 6 and Fig. 7a demonstrate that the proposed P-RS matcher is able to outperform the D-RS method and achieved comparable accuracy to FaceVACS on near infrared sketches. The P-RS algorithm compares favorably to other methods that have been tested on the CASIA HFB NIR dataset with training and testing sets containing disjoint sets of subjects. For example, Lei and Li's kernel coupled spectral regression achieved a TAR of 76.95 percent at $\mathrm{FAR} = 0.1\%$ [12]. Similarly, the LSNA method by Liao et al. achieved a TAR of roughly 70 percent at $\mathrm{FAR} = 0.1\%$ [8]. By comparison, the P-RS method achieved an average TAR of 95.8 percent at $\mathrm{FAR} = 0.1\%$. While the previous methods trained on roughly 3,000 total NIR and VIS images from roughly 150 subjects, the proposed P-RS method was trained on 266 images from 133 subjects. However, the previous methods did test the accuracy on test sets with more images per subject, which may or may not have contained additional face variations (the single NIR and VIS image per subject that we used were selected at random). Regardless, the improved accuracy using a smaller training set of subjects clearly demonstrates the value of the proposed P-RS method.

The CMC results of matching thermal face images to standard face images are shown in Fig. 7b. P-RS is able to achieve an average Rank-1 accuracy of 46.7 percent. By comparison, it is observed that FaceVACS had a Rank-1 accuracy of 21.5 percent. This drastic improvement demonstrates the benefit of P-RS's notable property of not requiring a feature descriptor that is invariant to changes in the probe and gallery modalities. While a Rank-1 accuracy of 46.7 percent still leaves more to be desired, the examples in Fig. 1 show that even humans would have difficulty in this recognition task. The only previous method on thermal to visible matching achieved a Rank-1 accuracy of 50.06 percent, but it was evaluated on only 47 subjects in the gallery [17]. By contrast, the Rank-1 accuracy for the proposed P-RS method was computed using a gallery consisting of 10,333 subjects.

Rank-1 Accuracy (%)

Method	NIR	Thermal	Sketch	Forensic*	Standard
P-RS	87.8 ± 4.53	46.7 ± 2.41	74.6 ± 5.42	14.7 ± 1.69	92.5 ± 1.91
D-RS	66.6 ± 6.97	41.5 ± 0.98	96.4 ± 1.54	17.4 ± 3.10	93.7 ± 0.20
(P-RS)+(D-RS)	86.6 ± 4.35	49.2 ± 1.90	92.5 ± 3.52	20.8 ± 2.07	93.0 ± 1.05
FaceVACS	87.8 ± 4.14	21.5 ± 0.83	84.8 ± 2.05	1.9 ± 1.03	98.7 ± 0.40

** Results for forensic sketch are the Rank-50 accuracy.*

(a)

TAR @ FAR = 1.0%

Method	NIR	Thermal	Sketch	Forensic	Standard
P-RS	98.2 ± 1.63	76.4 ± 2.55	99.5 ± 0.35	14.7 ± 3.38	96.8 ± 0.52
D-RS	94.0 ± 3.50	77.5 ± 1.22	99.6 ± 0.41	17.7 ± 6.88	97.9 ± 0.34
(P-RS)+(D-RS)	97.0 ± 2.36	78.2 ± 0.13	99.7 ± 0.27	18.9 ± 2.31	97.6 ± 0.34
FaceVACS	93.7 ± 1.63	47.5 ± 2.49	92.2 ± 1.50	2.6 ± 1.03	99.5 ± 0.40

(b)

TAR @ FAR = 0.1%

Method	NIR	Thermal	Sketch	Forensic	Standard
P-RS	95.8 ± 6.15	71.2 ± 12.94	99.0 ± 1.25	12.9 ± 5.39	95.5 ± 2.55
D-RS	90.8 ± 8.52	72.5 ± 12.25	99.5 ± 0.44	15.7 ± 7.89	96.7 ± 2.41
(P-RS)+(D-RS)	94.5 ± 6.45	72.7 ± 13.47	99.4 ± 0.73	16.0 ± 7.23	96.4 ± 2.41
FaceVACS	92.0 ± 4.39	44.4 ± 7.85	89.6 ± 6.45	2.5 ± 0.97	99.1 ± 1.02

(c)

Fig. 6. Recognition results for the proposed prototype random subspace and the baseline direct random subspace method across five recognition scenarios. Listed are the (a) rank retrieval results with using an additional 10,000 subjects in the gallery, (b) true accept rates at a fixed false accept rate (FAR) of 1.0 percent, and (c) TAR at FAR of 0.1 percent. (P-RS)+(D-RS) is a sum of score fusion between the two methods.

The CMC results of matching viewed sketch face images to standard face images are shown in Fig. 7c. The P-RS exhibited decreased accuracy when using the extended gallery. However, the TAR values in Figs. 6b and 6c indicate that P-RS is still well suited for this task. When removing the extended gallery, the P-RS method achieved an average Rank-1 recognition rate of 99.47 percent. This is slightly lower than the 99.87 percent average accuracy reported in [12]. Using P-RS parameters tuned to the viewed sketch scenario (instead of the generic parameters used) should improve this accuracy.

The CMC results of matching forensic sketch face images to standard face images are shown in Fig. 7d. For forensic sketches, the Rank-50 accuracy is most relevant because the Rank-1 accuracy is too low to be useful: Forensic investigators generally examine roughly the top 50 retrieved matches from a query. In this forensic sketch scenario, the P-RS method (Rank-50 accuracy of 14.7 percent) was outperformed by the D-RS method (Rank-50 accuracy of 17.4 percent). The only previous method to publish results on forensic sketch matching also used the same size extended gallery and achieved a Rank-50 accuracy of 13.4 percent [9] (this number is the weighted average of a 32.65 percent Rank-50 accuracy on 49 *good* sketches and an 8.16 percent accuracy on 110 *poor* sketches). It is important to note that the matcher in [9] was trained on viewed sketches and not forensic sketches like P-RS and D-RS.

The lower accuracy of P-RS compared to D-RS on the forensic sketch dataset can be attributed to two factors. The primary factor is the small training set. While both methods utilize learning, D-RS is able to leverage the a priori knowledge that SIFT and MLBP perform well for direct similarity measurement. An additional reason for P-RS's lower accuracy on forensic sketch matching is that these sketches are often not completely accurate due to the inability of a witness to adequately describe the face of a suspect. This impacts the desired property in (8) that if two images of the same subject are similar in the probe modality that they will be similar in the gallery modality. Despite these limitations, P-RS still achieved an order of magnitude improvement in accuracy over the COTS FRS.

Fig. 9 demonstrates the ability of the P-RS framework to perform recognition using different feature descriptors for the probe and gallery images. Fig. 9 lists the accuracy for the viewed sketch and near-infrared scenarios averaged over five random training/testing splits without the additional 10,000 gallery images. The columns indicate each of the six different image filter and feature descriptor combinations used to represent the gallery, and the rows indicate the representations used for the probe images. Thus, the nondiagonal entries for each scenario correspond to when the probe and gallery images are represented with different features. The accuracy is generally higher when the same features are used for faces in the probe and gallery (i.e., the diagonal entries). Various levels of accuracy are witnessed when using different image features, ranging from poor to high. Similar results were obtained with the

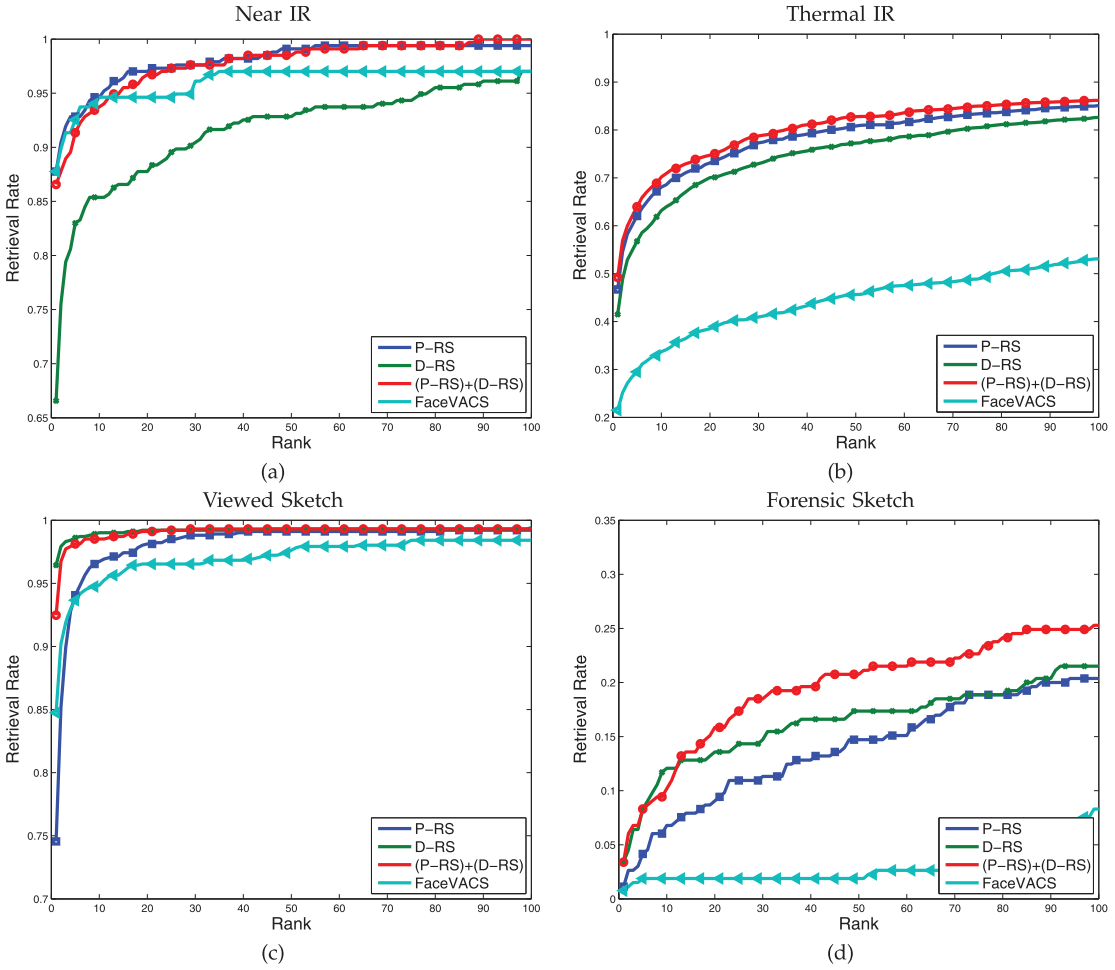


Fig. 7. CMC plots for each of the four HFR scenarios tested. Each scenario is using an additional 10,000 gallery images to better replicate real-world matching scenarios. Listed are the accuracies for the proposed prototype random subspace method, the direct random subspace method [10], the sum-score fusion of P-RS and D-RS, and Congitec's FaceVACS system [28].

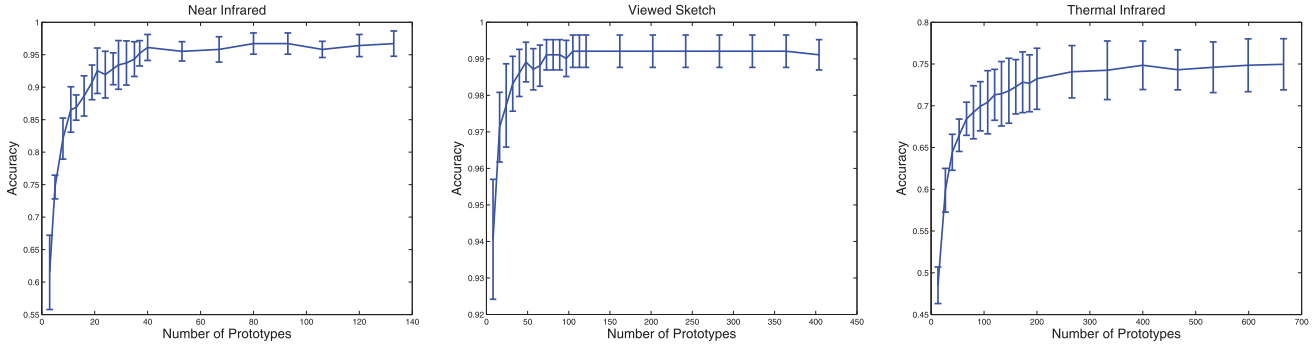
thermal and forensic sketch scenarios, but were omitted due to limited space.

The ability to perform face recognition with the probe and gallery images using different representations is a property that previous feature-based methods did not possess. This property is important to mention because it demonstrates the proposed method's ability to generalize to other unknown HFR scenarios. For example, in the case of thermal to visible recognition, if a local feature descriptor is developed that leads to high accuracy in matching thermal to thermal, it can be incorporated into this framework even if it does not work well in the visible domain. As other HFR scenarios are attempted (such as 3D to 2D visible photograph), this property could prove extremely useful in overcoming the hurdle of finding a feature descriptor that is invariant to changes between the two domains which feature-based methods rely on.

The relationship between the number of prototypes used in the P-RS representation and the recognition accuracy was measured to better understand the number of prototypes needed for maximal performance. Fig. 8 lists the recognition accuracy when all the training subjects are available for LDA subspace learning, but only a subset of the subjects are used as prototypes. As shown, the recognition accuracy generally saturates around 100 prototypes. The scenarios with more subjects available for LDA (e.g., thermal) do seem to benefit from more prototypes than those with fewer LDA subjects (e.g., NIR). However, in all cases the number of prototypes needed is significantly less than the total number of training subjects. Such results speak well toward the scalability of the proposed P-RS method.

The value of both the P-RS and D-RS methods is observed when examining their fusion results. In all cases except near infrared HFR, the fusion of the two methods resulted in improved recognition accuracy. This indicates that the two methods are generally able to extract some degree of orthogonal information, resulting in improved accuracy through fusion.

The proposed P-RS framework also generalizes to standard face recognition scenarios. Using the standard (nonheterogeneous) face dataset, Fig. 10a compares the accuracy of P-RS, D-RS, and FaceVACS. FaceVACS clearly outperforms P-RS and D-RS as expected since it is consistently one of the top performers in NIST face recognition benchmarks. However, using four different face datasets we see that P-RS and D-RS both achieve Rank-1 accuracies around 93 percent with 10,876 subjects in the gallery. In Fig. 10b, the results of matching using different feature descriptors in the probe and gallery domain are shown. The ability to match probe and gallery images using



(a) (b) (c)

Fig. 8. The effect of the number of prototypes on the recognition accuracy. The accuracy (vertical axis) is the TAR @ $\mathrm{FAR} = 1.0\%$. For each HFR scenario, results are computed using a subset of the prototypes (horizontal axis) for the initial representation, and all training samples for the subsequent LDA step. For each scenario, it is observed that accuracy generally plateaus around 100 prototypes. However, when more LDA training samples are available (e.g., thermal), additional prototypes seem to help the recognition accuracy. Results shown are the mean and standard deviation from the five random splits for the following HFR scenarios: (a) near infrared, (b) viewed sketch, and (c) thermal infrared.

different feature representations is rare and could benefit situations in which older face templates exist but the original face images are missing.

The proposed P-RS method is computationally scalable to meet the demands of real-world face recognition systems. Running in Matlab and using a single core from a 2.8 GHz Intel Xeon processor, the following amount of time is needed to enroll a single face image after the offline training step. Image filtering requires roughly 0.008 sec for DoG, 1.1 sec for CSDN, and 0.004 sec for Gauss. The MLBP and SIFT feature descriptors each take roughly 0.35 sec to compute. Because each image filtering is performed only once, and each feature descriptor is computed three times (once for each filter), computing all six filter/descriptor combinations takes around 3.2 sec. The prototype random subspace representation with 30 bags takes roughly 0.2 sec to compute for a single filter/descriptor combination. Thus, all six filter/descriptor combinations take roughly 1.8 sec. In total, a face image needs around 5.0(s) to enroll in Matlab. With a gallery of n_g subjects and the final feature vector Φ of size d', identification of a subject is $O(d' \cdot n_g)$. Depending on the number of bags, the number of prototypes for each scenario, and the variance retained in the PCA step, d' is in the order of 1,000.

8 Summary

A method for heterogeneous face recognition, called Prototype Random Subspaces, is proposed. Probe and gallery images are initially filtered with three different image filters, and two different local feature descriptors are then extracted. A training set acts as a set of prototypes in which each prototype subject has an image in both the gallery and probe modalities. The nonlinear kernel similarity between an image and the prototypes is measured in the corresponding modality. A random subspace framework is employed in conjunction with LDA subspace analysis to further improve the recognition accuracy.

The proposed method leads to excellent matching accuracies across four different HFR scenarios (near infrared, thermal infrared, viewed sketch, and forensic sketch). Results were compared against a leading commercial face recognition engine. In most of our experiments the gallery size was increased with an additional 10,000 subjects to better replicate real matching scenarios. In addition to excellent matching accuracies, one key benefit of the proposed P-RS method is that different feature descriptors can be used to represent the probe and gallery images. Finally, the number of prototypes needed by P-RS for effective performance was shown to be stable.

Viewed Sketch

Probe Features	Gallery Features: DoG SIFT	DoG MLBP	CSDN SIFT	CSDN SIFT	Gauss SIFT	Gauss MLBP
DoG SIFT	**98.5**	99.0	98.0	99.0	98.5	99.0
DoG MLBP	94.1	**95.0**	92.1	93.6	94.6	95.5
CSDN SIFT	98.5	99.0	**99.0**	99.0	98.5	99.0
CSDN SIFT	94.1	83.7	88.1	**96.0**	91.6	93.1
Gauss SIFT	97.5	98.0	98.5	98.5	**99.0**	98.0
Gauss MLBP	90.6	84.2	94.6	94.1	95.5	**98.5**

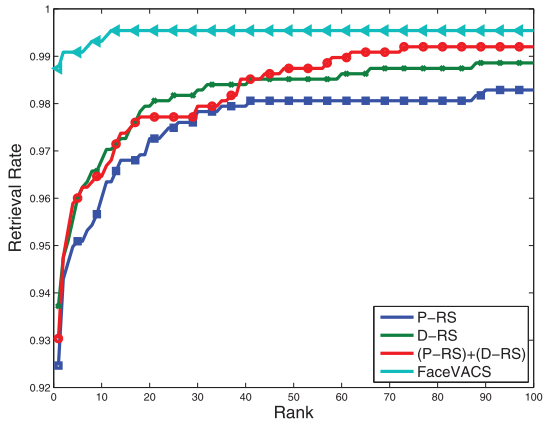
(a)

Near Infrared

Probe Features	Gallery Features: DoG SIFT	DoG MLBP	CSDN SIFT	CSDN SIFT	Gauss SIFT	Gauss MLBP
DoG SIFT	**94.0**	92.5	92.5	92.5	92.5	91.0
DoG MLBP	83.6	**91.0**	74.6	80.6	83.6	82.1
CSDN SIFT	94.0	92.5	**97.0**	88.1	94.0	89.6
CSDN SIFT	92.5	89.6	94.0	**98.5**	89.6	91.0
Gauss SIFT	86.6	85.1	91.0	82.1	**92.5**	82.1
Gauss MLBP	68.7	74.6	76.1	76.1	88.1	**89.6**

(b)

Fig. 9. True accept rates at a fixed false accept rate of 1.0 percent using the proposed P-RS framework with different features for the probe and gallery. The rows list the features used to represent the probe images, and the columns list the features for the gallery images. The nondiagonal entries in each table (in bold) use different feature descriptor representations for the probe images than the gallery images. These results demonstrate another "heterogeneous" aspect of the proposed framework: recognition using heterogeneous features. The first row and sixth column of (a) demonstrate that the P-RS framework can achieve 99.0 percent accuracy representing the probe/sketch images with DoG filtered SIFT and the photo/gallery images with Gaussian filtered MLBP.



(a)

Probe Features	Gallery Features: DoG SIFT	DoG MLBP	CSDN SIFT	CSDN SIFT	Gauss SIFT	Gauss MLBP
DoG SIFT	**97.9**	96.9	97.6	96.9	96.9	97.3
DoG MLBP	97.6	**96.9**	97.3	97.3	96.2	96.6
CSDN SIFT	95.5	95.5	**96.6**	95.5	95.2	95.5
CSDN SIFT	95.9	94.5	96.2	**95.9**	96.2	95.9
Gauss SIFT	95.9	95.2	95.5	95.5	**96.6**	95.2
Gauss MLBP	95.9	96.2	95.2	95.5	96.2	**96.9**

(b)

Fig. 10. Face recognition results when using photographs for both the probe and gallery (i.e., nonheterogeneous face recognition). (a) CMC plot of matcher accuracies with an additional 10,000 gallery images. (b) Results when different features are used to represent the probe and gallery images. The layout is the same as in Fig. 9.

Future work will focus on 1) improving the accuracy of each of the tested HFR scenarios separately, and 2) incorporating additional HFR scenarios. Tailoring the P-RS parameters and learning weighted fusion schemes for each HFR scenario separately should offer further accuracy improvements. One additional HFR scenario that should be considered is 3D to 2D face matching. P-RS should be particularly impactful in this scenario because heterogeneous features will be required to represent faces in the 3D and 2D modalities.

Acknowledgments

The authors would like to thank Scott McCallum and the rest of the his team at the Pinellas County Sheriff's Office, and Captain Greg Michaud from the Michigan State Police for their gracious support of this research. They would also like to thank Rong Jin and Serhat Bucak for their feedback on this research. This manuscript benefited from the value observations provided in the review process. Anil Jain's research was partially supported by the World Class University (WCU) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

References

[1] S. Li, *Encyclopedia of Biometrics.* Springer, 2009.
[2] P.J. Grother, G.W. Quinn, and P.J. Phillips, "MBE 2010: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms," *Nat'l Inst. of Standards and Technology*, vol. 7709, 2010.
[3] X. Tang and X. Wang, "Face Sketch Recognition," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 50-57, Jan. 2004.
[4] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A Nonlinear Approach for Face Sketch Synthesis and Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1005-1010, 2005.
[5] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955-1967, Nov. 2009.
[6] X. Gao, J. Zhong, J. Li, and C. Tian, "Face Sketch Synthesis Algorithm Based on E-HMM and Selective Ensemble," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 487-496, Apr. 2008.
[7] W. Zhang, X. Wang, and X. Tang, "Lighting and Pose Robust Face Sketch Synthesis," *Proc. European Conf. Computer Vision*, 2010.
[8] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li, "Heterogeneous Face Recognition from Local Structures of Normalized Appearance," *Proc. Third Int'l Conf. Advances in Biometrics*, 2009.
[9] B. Klare, Z. Li, and A. Jain, "Matching Forensic Sketches to Mugshot Photos," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639-646, Mar. 2011.
[10] B. Klare and A. Jain, "Heterogeneous Face Recognition: Matching NIR to Visible Light Images," *Proc. Int'l Conf. Pattern Recognition*, 2010.
[11] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "On Matching Sketches with Digital Face Images," *Proc. IEEE Conf. Biometrics: Theory, Applications, and Systems*, pp. 1-7, 2010.
[12] Z. Lei and S. Li, "Coupled Spectral Regression for Matching Heterogeneous Faces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1123-1128, 2009.
[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, July 2002.
[14] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
[15] D. Yi, S. Liao, Z. Lei, J. Sang, and S. Li, "Partial Face Matching between Near Infrared and Visual Images in MBGC Portal Challenge," *Proc. Third Int'l Conf. Advances in Biometrics*, pp. 733-742, 2009.
[16] W. Zhang, X. Wang, and X. Tang, "Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[17] J. Li, P. Hao, C. Zhang, and M. Dou, "Hallucinating Faces from Thermal Infrared Images," *Proc. Int'l Conf. Image Processing*, pp. 465-468, 2008.
[18] D. Lin and X. Tang, "Inter-Modality Face Recognition," *Proc. European Conf. Computer Vision*, 2006.
[19] M.-F. Balcan, A. Blum, and S. Vempala, "Kernels as Features: On Kernels, Margins, and Low-Dimensional Mappings," *Machine Learning*, vol. 65, pp. 79-94, 2006.
[20] A. Quattoni, M. Collins, and T. Darrell, "Transfer Learning for Image Classification with Sparse Prototype Representations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
[21] K.I. Kim, K. Jung, and H.J. Kim, "Face Recognition Using Kernel Principal Component Analysis," *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40-42, Feb. 2002.
[22] Q. Liu, H. Lu, and S. Ma, "Improving Kernel Fisher Discriminant Analysis for Face Recognition," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42-49, Jan. 2004.
[23] L. Juwei, K. Plataniotis, and A. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 117-126, Jan. 2003.
[24] A. Gyaourova and A. Ross, "Index Codes for Multibiometric Pattern Retrieval," *IEEE Trans. Information Forensics and Security*, vol. 7, no. 2, pp. 518-529, Apr. 2012.
[25] B. Klare and A. Jain, "Sketch to Photo Matching: A Feature-Based Approach," *Proc. SPIE Conf. Biometric Technology for Human Identification*, 2010.
[26] M. Riesenhuber and T. Poggio, "Hierarchical Models of Object Recognition in Cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019-1025, 1999.
[27] B. Klare and A.K. Jain, "On a Taxonomy of Facial Features," *Proc. IEEE Conf. Biometrics: Theory, Applications, and Systems*, 2010.
[28] FaceVACS Software Developer Kit, Cognitec Systems GmbH, http://www.cognitec-systems.de, 2012.

[29] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions," *IEEE Trans. Image Processing,* vol. 19, no. 6, pp. 1635-1650, June 2010.
[30] E. Meyers and L. Wolf, "Using Biologically Inspired Features for Face Processing," *Int'l J. Computer Vision,* vol. 76, no. 1, pp. 93-104, 2008.
[31] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 2037-2041, Dec. 2006.
[32] S. Li, R. Chu, S. Liao, and L. Zhang, "Illumination Invariant Face Recognition Using Near-Infrared Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 4, pp. 627-639, Apr. 2007.
[33] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.
[34] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.
[35] C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. Image Processing,* vol. 11, no. 4, pp. 467-476, Apr. 2002.
[36] S. Raudys and A. Jain, "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 3, pp. 252-264, Mar. 1991.
[37] H.Y. Jie, H. Yu, and J. Yang, "A Direct LDA Algorithm for High-Dimensional Data—with Application to Face Recognition," *Pattern Recognition,* vol. 34, pp. 2067-2070, 2001.
[38] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition," *Pattern Recognition Letters,* vol. 26, no. 2, pp. 181-191, 2005.
[39] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832-844, Aug. 1998.
[40] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996.
[41] T. Dietterich, "Ensemble Methods in Machine Learning," *Multiple Classifier Systems,* vol. 1857, pp. 1-15, 2000.
[42] X. Wang and X. Tang, "Random Sampling for Subspace Face Recognition," *Int'l J. Computer Vision,* vol. 70, no. 1, pp. 91-104, 2006.
[43] A. Martinez and R. Benavente, "The AR Face Database," Technical Report 24, CVC, 1998.
[44] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The Extended M2VTS Database," *Proc. Audio and Video-Based Biometric Person Authentication,* 1999.
[45] L. Gibson, *Forensic Art Essentials.* Elsevier, 2008.
[46] K. Taylor, *Forensic Art and Illustration.* CRC Press, 2001.
[47] P. Phillips, H. Moon, P. Rauss, and S. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1997.

Brendan F. Klare received the BS and MS degrees in computer science from the University of South Florida in 2007 and 2008, respectively, and the PhD degree in computer science from Michigan State University in 2012. He is a lead scientist at Noblis. From 2001 to 2005, he served as an airborne ranger infantryman in the 75th Ranger Regiment. His research interests include pattern recognition, image processing, and computer vision. He has authored several papers on the topic of face recognition, and he received the Honeywell Best Student Paper Award at the 2010 IEEE Conference on Biometrics: Theory, Applications, and Systems (BTAS). He is a member of the IEEE.

Anil K. Jain is a University Distinguished Professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the *IEEE Transaction on Pattern Analysis and Machine Intelligence* (1991-1994). The holder of six patents in the area of fingerprints, he is the author of a number of books, including *Introduction to Biometrics* (2011), *Handbook of Face Recognition* (second edition, 2011), *Handbook of Fingerprint Recognition* (2009), *Handbook of Biometrics* (2007), *Handbook of Multibiometrics* (2006), *Biometrics: Personal Identification in Networked Society* (1999), and *Algorithms for Clustering Data* (1988). He served as a member of the Defense Science Board and The National Academies committees on "Whither Biometrics and Improvised Explosive Devices." He received the 1996 *IEEE Transactions on Neural Networks* Outstanding Paper Award and the Pattern Recognition Society best paper awards in 1987, 1991, and 2005. He has received Fulbright, Guggenheim, Alexander von Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu awards. ISI has designated him a highly cited researcher. According to Citeseer, his book *Algorithms for Clustering Data* (Prentice-Hall, 1988) is ranked number 93 in most cited articles in computer science. He is a fellow of the IEEE, AAAS, ACM, IAPR, and SPIE.