Mutual Component Analysis for Heterogeneous Face Recognition

ZHIFENG LI, Chinese Academy of Sciences
DIHONG GONG, Chinese Academy of Sciences
QIANG LI, University of Technology Sydney
DACHENG TAO, University of Technology Sydney
XUELONG LI, Chinese Academy of Sciences

Heterogeneous face recognition, also known as cross-modality face recognition or inter-modality face recognition, refers to matching two face images from alternative image modalities. Since face images from different image modalities of the same person are associated with the same face object, there should be mutual components that reflect those intrinsic face characteristics that are invariant to the image modalities. Motivated by this rationality, we propose a novel approach called mutual component analysis (MCA) to infer the mutual components for robust heterogeneous face recognition. In the MCA approach, a generative model is first proposed to model the process of generating face images in different modalities, and then an Expectation Maximization (EM) algorithm is designed to iteratively learn the model parameters. The learned generative model is able to infer the mutual components (which we call the *hidden factor*, where *hidden* means the factor is unreachable and invisible, and can only be inferred from observations) that are associated with the person's identity, thus enabling fast and effective matching for cross-modality face recognition. To enhance recognition performance, we propose an MCA-based multi-classifier framework using multiple local features. Experimental results show that our new approach significantly outperforms the state-of-the-art results on two typical application scenarios, sketch-to-photo and infrared-to-visible face recognition.

Categories and Subject Descriptors: I.5.1 [**Pattern Recognition**]: Models

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Face recognition, heterogeneous face recognition, mutual component analysis (MCA)

ACM Reference Format:
Zhifeng Li, Dihong Dong, Qiang Li, Dacheng Tao, and Xuelong Li, 2015. Mutual Component Analysis for Heterogeneous Face Recognition *ACM Trans. Intell. Syst. Technol.* 9, 4, Article 39 (July 2015), 22 pages.
DOI: http://dx.doi.org/10.1145/2807705

This work was supported by grants from National Natural Science Foundation of China (61103164 and 61125106), Natural Science Foundation of Guangdong Province (2014A030313688), Australian Research Council Projects (FT-130101457 and LP-140100569), Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Guangdong Innovative Research Team Program (No.201001D0104648280), the Key Research Program of the Chinese Academy of Sciences (Grant No. KGZD-EW-T03), and project MMT-8115038 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.
Author's addresses: Z. Li and D. Gong, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, P. R. China; e-mail: {zhifeng.li, dh.gong}@siat.ac.cn; Q. Li and D. Tao, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway, Ultimo, NSW 2007, Australia; e-mail: qiang.li-2@student.uts.edu.au, dacheng.tao@uts.edu.au; X. Li, the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China; e-mail: xuelong_li@opt.ac.cn.

DOI: http://dx.doi.org/10.1145/2807705

1. INTRODUCTION

Face recognition has attracted great attention in recent years due to the increasing demands of real-world applications [Chellappa et al. 1995; Zhao et al. 2003; Wolf et al. 2008; Wright and Hua 2009; Li and Tang 2007; Li 2009; Tang and Li 2009; Berretti et al. 2012; Ewerth et al. 2012; Jia and Gong 2008; Zhang et al. 2011b; Li et al. 2011; Li et al. 2009; Shao and Fu 2013; Ding et al. 2015; Li et al. 2015; Li et al. 2014; Shao et al. 2014a; Shao et al. 2014b; Zhang et al. 2014; Yan et al. 2013; Zhen et al. 2014; Li et al. 2014]. An emerging research topic in the face recognition community is heterogeneous face recognition [Li 2009; Ding et al. 2014; Shao et al. 2015]. Heterogeneous face recognition, also known as inter-modality face recognition or cross-modality face recognition, refers to matching two face images from alternative image modalities. To illustrate the cross-modality face recognition problem, two typical cases are considered: sketch-photo recognition and infrared-visible recognition. In the first case, since sketches are drawn by artists, they look quite different from the corresponding photos, as shown in Fig. 1(a). We can see that much texture information is missing from the sketches, and outlines are distorted. In the second case, we need to match images taken by an infrared device to gallery images taken by an visible-light camera [Yi et al. 2007; Liao et al. 2009; Klare and Jain 2010; Li et al. 2014], as shown in Fig. 1(b). Infrared-visible face recognition confronts similar problems when query images of objects are taken in a dark environment where infrared cameras are employed to deal with the weak illumination problem. The challenge of matching heterogeneous face images lies in the fact that the probe images and the gallery images are acquired through different processes under different conditions and thus have significant discrepancies.

A large number of studies on the heterogeneous face recognition problem have been conducted which essentially fall into the following three categories: 1) Convert images from one modality to the other by synthesizing a pseudo-image from the query image such that the matching process can be carried out within the same modality. For instance, [Tang and Wang 2004] applied a holistic mapping to convert a photo image into a corresponding sketch image. In [Liu et al. 2005; Wang and Tang 2009; Geng and Jiang 2011], the authors used local patch-based mappings to convert images from one modality to the other for sketch-photo recognition. 2) Design an appropriate representation that is insensitive to the modalities of images. For example, Klare et al [Klare et al. 2011] used SIFT [Lowe 2004] feature descriptors and multi-scale local binary patterns [Ojala et al. 2002] to represent both the sketch and photo images; Zhang et al [Zhang et al. 2011a] proposed a learning-based algorithm to capture discriminative local face structures and effectively match photo and sketch. 3) Compare the heterogeneous images on a common subspace where the modality difference is believed to be minimized. Tenenbaum et al. [Tenenbaum and Freeman 2000] applied the Bilinear Model (BLM) by Singular Value Decomposition (SVD) to develop a common content space (associated with identity) for a set of different styles (corresponding to modalities). In [Lin and Tang 2006], Lin et al. developed a common discriminant feature extraction (CDFE) algorithm for the inter-modality face recognition problem. In [Yi et al. 2007], Yi et al used the Canonical Correlation Analysis (CCA) technique to construct a common subspace in which the correlations between infrared and visible images can be maximized. In [Li et al. 2009], Li et al applied the CCA on cross-pose face recognition. In [Sharma and Jacobs 2011], Sharma et al applied the Partial Least Squares (PLS) method to derive a linear subspace in which cross-modality images are highly correlated, while variances are better preserved at the same time compared to the previous CCA method. In [Huang et al. 2013], Huang et al. proposed an effective subspace learning framework called Regularized Discriminative Spectral Regression for heterogeneous face recognition. In [Klare and Jain 2013], a generic HFR frame-

(a) (b)

Fig. 1. Example images for: (a) sketches (top row) and photos (bottom row); (b) visible (top row) and infrared (bottom row) faces.

work was proposed in which both probe and gallery images are represented in terms of non-linear kernel similarities to a collection of prototype face images to enhance heterogeneous face recognition accuracy.

In this paper, we propose a novel subspace approach called mutual component analysis (MCA) for cross-modality face recognition. Our primary idea is that face images of the same person should be associated with the same mutual component (which we call the *hidden factor*, where *hidden* means the factor is unreachable and invisible, and can only be inferred from observations) that is invariant to modalities, as illustrated by Figure 2. By learning a generative model to model the generation process of face images in different modalities, the cross-modality face recognition can be easily accomplished by inferring the hidden factors that are associated with the person's identity while invariant to its image modality. Our phenomenon-to-nature idea is quite suitable for this cross-modality problem since we can capture the unchanged nature (modeled by the hidden factors) of different phenomenon (modalities), based on which the cross-modality face recognition problem can be well solved, as supported by our experimental results.

The major contributions of this paper are summarized as follows:

(1) Considering the specific property of cross-modality face recognition, we propose a novel subspace approach called mutual component analysis (MCA) to address this challenging problem. Unlike the previous methods in heterogeneous face recognition, our approach provides new insight into this problem by constructing a generative model to formulate the generation of facial images in different modalities. By using this generative model, the mutual components (which are only associated with the persons identity and invariant to the image modality) can be extracted for effective classification. To the best of our knowledge, this is a novel method that has not previously been attempted in this field.
(2) Based on the proposed MCA approach, we develop an enhancement called the MCA-based multi-classifier fusion framework, using multiple local features to improve recognition performance.
(3) From the experiments, we can clearly see that our proposed approach is able to obtain significant improvement over published algorithms in the literature on two large and challenging cross-modality face recognition databases.

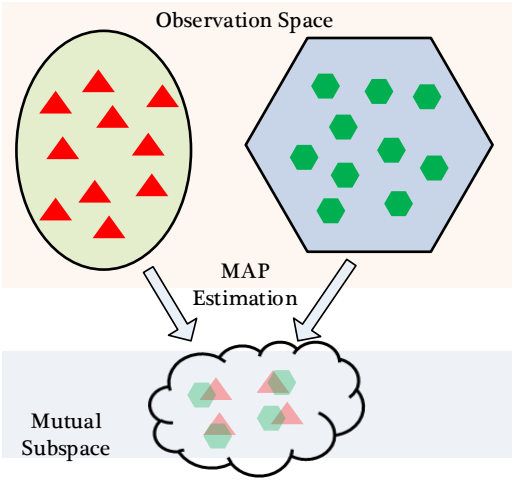

Fig. 2. Illustration for inferring the hidden factor from observation.

The rest of this paper is organized as follows: In Section II, we describe the proposed mutual component analysis (MCA) approach for cross-modality face recognition. In Section III, an MCA-based multi-classifier fusion framework is proposed to further enhance recognition performance. Experimental results and analysis are given in Section IV. Lastly, Section V concludes this paper.

2. MUTUAL COMPONENT ANALYSIS

In this section, we first propose the linear generative model which we name mutual component analysis (MCA) for the cross-modality problem. Then an EM algorithm is developed to learn the model parameters. Lastly, an MCA-based linear classifier is constructed for cross-modality face recognition.

2.1. Linear Generative Model

By viewing the cross-modality face recognition problem as the observation of features through different transmission channels, we can formulate it with a generative model. In this sense, the face data of different modalities can be considered as being generated by the same original signal (associated with the mutual components) passing through different transmission channels.

To simplify the analysis, we introduce (or re-define) the following notation, where a term with superscript k indicates that it is dependent on modality k, and a term with subscript i indicates that it is dependent on the i-th observation.

N: The number of training observations for each modality.
D: The dimension of observations (the dimension of features of images in our experiments).
d: The dimension of hidden factor $\vec{y}$ (d is decided empirically).
r: The rank of complemental matrix U^k (r is decided empirically).
K: The number of modalities ($K = 2$ in our sketch-photo/infrared-visible face recognition).
M^k: $D \times N$ matrix with N training observations (columns) in modality k.
I_d: The $d \times d$ identity matrix.
$\vec{m}_i^k$: The i-th $D \times 1$ column vector in the training set M^k.

M_i: The i-th observations in all modalities $M_i = \{\vec{m}_i^k | k = 1, , K\}$, K is the number of modalities.
$\vec{y}_i$: The $d \times 1$ vector generates the observations M_i, with prior standard normal distribution.
Y: $d \times N$ matrix with the i-th column consisting of $\vec{y}_i$.
Σ^k: $D \times D$ diagonal matrix ($\Sigma^k = C^k {C^k}^T$, we only reserve the diagonal components for computational efficiency).
$\vec{\beta}^k$: The estimated $D \times 1$ mean vector for observations in modality k.
U^k: The $D \times d$ matrix of rank at most r ($U^k = R^k * W$, by adapting the matrix R^k, we gain a rank-constrained U^k. W is a $r \times d$ constant orthonormal matrix).

Suppose that we have two sets of observations (where $K = 2$ in this case) $M^1 = \{\vec{m}_1^1, \vec{m}_2^1, \ldots, \vec{m}_N^1\}$ and $M^2 = \{\vec{m}_1^2, \vec{m}_2^2, \ldots, \vec{m}_N^2\}$ of N pairs of subjects consisting of $D \times 1$ column vectors, where the superscripts indicate the modalities of observations and the subscripts indicate the identities. Since M^1 and M^2 are two observations of the same group of subjects, they should be associated with the same set of mutual components $H = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}$ consisting of the $d \times 1$ column vectors that generate them.

Since observations in both M^1 and M^2 are generated by the corresponding mutual components in H, the relationship between H and M can be modeled by a transmission channel denoted by a transmission function F. For the k^{th} modality, there is a function from d-space to D-space $\vec{F}^k : R^{d \times 1} \to R^{D \times 1}$:

$$\vec{F}^k(\vec{x}) = [\vec{F}_1^k(\vec{x}), \vec{F}_2^k(\vec{x}), \ldots, \vec{F}_D^k(\vec{x})]^T, \tag{1}$$

such that

$$\vec{m}_i^k = F^k(\vec{h}_i). \tag{2}$$

We assume that H is Gaussian distributed: $H \sim N(\vec{\mu}, \Sigma)$, which is equivalent to:

$$\vec{h}_i = \vec{\mu} + V\vec{y}_i, \tag{3}$$

where $\Sigma = VV^T$, $\vec{y} \sim N(\vec{0}, I)$. Substituting (3) into (2), we have:

$$\vec{m}_i^k = F^k(\vec{\mu} + V\vec{y}_i).$$

where: $D \times 1$ vector $\vec{\beta}^k = \vec{F}^k(\vec{\mu})$, and:

$$\vec{x}_i = V\vec{y}_i. \tag{4}$$

$J_k(\vec{\mu})$ is the $D \times d$ Jacobian matrix with element $J_k(p, q) = \left.\frac{\partial F_p^k(\vec{x})}{\partial x_q}\right|_{\vec{x}=\vec{\mu}}$, representing the partial derivative of the p-th component function in (1) with respect to the q-th variable of $\vec{x}$. $D \times 1$ vector $\vec{r}_i^k$ is the high-order residue with the j-th element $\vec{r}_i^k(j)$. Here Φ_j^k is the $d \times d$ Hessian matrix with element $\Phi_j^k(p, q) = \left.\frac{\partial^2 F_j^k(\vec{x})}{\partial x_p \partial x_q}\right|_{\vec{x}=\vec{\mu}}$,

To simplify our model, we use Gaussian random noise to model the high-order derivatives of the function $\vec{F}^k$. Thus our linear generative model can finally be formulated as:

$$\vec{m}_i^k = \vec{\beta}^k + T^k\vec{y}_i + C^k\vec{z}, \tag{5}$$

where $T^k = J_k(\vec{\mu})V$ is a $D \times d$ matrix, C^k is a $D \times D$ matrix with each column representing the bases of the noises (corresponding to the eigenvectors of the covariance matrix of noises), and $\vec{z}$ is the $D \times 1$ random vector whose components are independent standard normal random variables.

According to the linear generative model given by (5), given the modality-dependent observation $\vec{m}_i^k$, our aim is to infer the *posterior* distribution $P(\vec{y}_i|\vec{m}_i^k)$ of the hidden factor $\vec{y}_i$ (we refer to $\vec{y}_i$ as the hidden factor while $\vec{h}_i$ is the mutual component in the rest of this paper), based on which modality-independent face recognition can be implemented, as illustrated by Fig. 3.

2.2. Model Parameter Estimation

In this subsection, we construct a Maximum A Posteriori (MAP)-based EM algorithm to iteratively estimate the model parameters $\theta^k = \{\vec{\beta}^k, T^k, \Sigma^k\}$ in equation (5) (we will use $\Sigma^k = C^k {C^k}^T$ to represent the C^k parameter in the rest of this paper).

The estimation of $\vec{\beta}^k$. Given two training sets of observations $M^1 = \{\vec{m}_1^1, \vec{m}_2^1, \ldots, \vec{m}_N^1\}$ and $M^2 = \{\vec{m}_1^2, \vec{m}_2^2, \ldots, \vec{m}_N^2\}$ of N subjects, the estimation of $\vec{\beta}^k$ is rather straightforward:

$$\vec{\beta}^k = \frac{1}{N}\sum_{i=1}^{N} \vec{m}_i^k. \tag{6}$$

The estimation of T^k *and* Σ^k. Since the features of face images are usually of high dimension, the estimation of T^k requires a large number of training samples to achieve sufficient accuracy, which may be practically infeasible in many situations. We deal with this problem by decomposing T^k into two parts:

$$T^k = S + U^k. \tag{7}$$

The S interprets the common components among different modalities, while $U^k = R^k * W$ is a complement for modality discrepancy (to constrain the rank of U^k to prevent over-fitting, we decompose the U^k into two matrices of rank r: R^k is a $D \times r$ ($r < D$) matrix and W is a $r \times d$ ($r < d$) constant orthonormal matrix where $WW^T = I$. By adapting R^k, we gain the matrix U^k of rank r, and the r is decided empirically. In our experiments, we found that the algorithm has a wide range for r to obtain the best performance). Thus the estimation of T^k can be divided into two steps: first estimate the common component S, and secondly estimate the modality-dependent component U^k. Substitute (7) into (5):

$$\vec{m}_i^k = \vec{\beta}^k + S\vec{y}_i + U^k\vec{y}_i + C^k\vec{z}. \tag{8}$$

In estimating the S, we first drop the relatively minor parts of $U^k\vec{y}_i + C^k\vec{z}$, and this approximation will be made up later by adapting U^k and C^k. Dropping the last two terms in (8), it yields:

$$\vec{m}_i^k = \vec{\beta}^k + S\vec{y}_i. \tag{9}$$

PROPOSITION 2.1. *Given K sets $M^1 \ldots M^K$ of observations (K is the number of modalities), and estimation of $\vec{\beta}^k$ in (6), the estimation of S in (9) is:*

$$S = V * \sqrt{E}. \tag{10}$$

where V is a $D \times d$ matrix whose columns consist of the eigenvectors of CM corresponding to the largest d eigenvalues (CM consists of the estimates of the covariance matrix of observations), and E is a $d \times d$ diagonal matrix whose diagonal elements consist of the corresponding eigenvalues.

The proof of Proposition 2.1 is given in the Appendix. In the rest of this subsection, we construct an EM algorithm to estimate the modality-dependent parameters U^k and C^k in (8).

Given an observation set M^k and a hidden factor matrix Y, our training aim is to maximize the joint probability:

$$\theta^k = \arg\max_{\theta} P_\theta(M^k, Y), \tag{11}$$

where $P_\theta(M^k, Y) = P(M^k, Y|\theta)$. However, the hidden factor matrix Y cannot be observed, and can only be inferred.

LEMMA 2.2. *Given model parameter estimate $\theta^k = (\vec{\beta}^k, T^k, \Sigma^k)$, the posterior distribution of hidden factor $\vec{y}_i$ is:*

$$P_{\theta^k}(\vec{y}_i|\vec{m}_i^k) \sim N(\vec{\varphi}^k, {L^k}^{-1}), \tag{12}$$

where $\vec{\varphi}^k = {L^k}^{-1}T^k{\Sigma^k}^{-1}(\vec{m}_i^k - \vec{\beta}^k)$ and $L^k = I_d + {T^k}^T{\Sigma^k}^{-1}T^k$ (I_d is $d \times d$ identity matrix).

The proof of Lemma 2.2 is given in the Appendix. Lemma 2.2 estimates the posterior distribution of hidden factors given observations. Instead of inferring one hidden factor per training sample, in our algorithm the samples with different modalities (from the same subject) share the same hidden factor. This is a major different between our approach and the traditional factor analysis based models. According to Lemma 2.2, the posterior distribution of hidden factor $\vec{y}_i$ is also Gaussian distributed and we can infer it with the model parameter θ^k. So far, we have seen that the estimation of parameter θ^k depends on the inference of $\vec{y}_i$; and the $\vec{y}_i$ again depends on the model parameters. To solve this co-dependent problem, we construct an EM algorithm to iteratively update the model parameter θ^k from a randomly initialized estimate θ^0. The following proposition describes how to update the R^k and Σ^k when given $\vec{\beta}^k$ and S.

PROPOSITION 2.3. *Given an initial estimate (R_0^k, Σ_0^k), let $\vec{E}^1_{M_i}$ and $\vec{E}^2_{M_i}$ denote the first and second moments of $\vec{y}_i$ calculated with the estimates on condition of $M_i = \{\vec{m}_i^k | k = 1, , K\}$. The new estimates (R^k, Σ^k) increase the joint probability of (11) if R^k is the solution of:*

$$\sum_{i=1}^{N}(\vec{m}_i^k - \vec{\beta}^k)\ {\vec{E}^1_{M_i}}^T = (S + R^k W)\sum_{i=1}^{N}\vec{E}^2_{M_i} \tag{13}$$

$$\Sigma^k = \frac{1}{N} diag\left\{\sum_{i=1}^{N}[\vec{m}_i^k - \vec{\beta}^k(S + R^k W)\vec{E}^1_{M_i}](\vec{m}_i^k - \vec{\beta}^k)^T\right\}. \tag{14}$$

The $diag\{\cdot\}$ means retaining the diagonal elements (we only retain the diagonal components for computational efficiency). The conditional moments are given by:

$$\vec{E}^1_{M_i} = \frac{1}{K}\sum_{k=1}^{K}\vec{E}^{1,k}_{M_i}, \tag{15}$$

$$\vec{E}^2_{M_i} = \frac{1}{K}\sum_{k=1}^{K}\left({L^k}^{-1} + \vec{E}^{1,k}_{M_i}\ {\vec{E}^{1,k}_{M_i}}^T\right), \tag{16}$$

where $\vec{E}^{1,k}_{M_i} = {L^k}^{-1}(S + R^k W)^T{\Sigma_0^k}^{-1}(\vec{m}_i^k - \vec{\beta}^k)$ represents the first moment for the hidden factor, and $L^k = I_d + (S + R^k W)^T{\Sigma^k}^{-1}(S + R^k W)$.

The proof of Proposition 2.3 is shown in the Appendix. Note that in Equation (15) and (16), the moments are computed as average moments of the estimates from different

modalities, which is different from the traditional factor analysis based models. This difference is not trivial because it essentially links the samples of different modalities together. Equation (13) is a linear regression system, and the least square solution provides the update rule:

$$R^k = \left\{ \sum_{i=1}^{N} (\vec{m}_i^k - \vec{\beta}^k)\ {\vec{E}_{M_i}^1}^T \left(\sum_{i=1}^{N} \vec{E}_{M_i}^2 \right)^+ - S \right\} W^T. \tag{17}$$

The detailed algorithm to estimate (T^k, Σ^k) is summarized as follows:

(a) Estimate S with equation (10).
(b) Initialize R^k and Σ^k.
(c) Compute the first and second moments $\vec{E}_{M_i}^1$ and $\vec{E}_{M_i}^2$ using (15) and (16).
(d) Update R^k and Σ^k using (17) and (14).
(e) Go to step (c) until convergence.
(f) Compute T^k using equation (7).

The proposed linear generative model distinguishes itself from the existing models in two major aspects. First, it is the first attempt to formulate the heterogeneous problem with hidden factors. Though there have been some existing works applying hidden factor analysis to solve computer vision problems, our algorithm is the first one to employ hidden factors to solve heterogeneous matching problem. Secondly, most of the existing works are using hidden factor analysis in an unsupervised manner, the proposed model learns hidden factors in a supervised manner. More specifically, the proposed model is able to learn a hidden factor that generates face images of different modalities for the same subject. This treatment allows the face images from the same subject of different modalities map to the same hidden factor, which is quite beneficial to the modality-invariant face recognition.

2.3. MCA-based Linear Classifier

In this subsection, we describe our linear classifier for heterogeneous face recognition based on the MCA model. Given the model parameter $\theta^k = \{\vec{\beta}^k, T^k, \Sigma^k\}$, according to (12), the posterior distribution of the hidden factors for any observation is Gaussian distributed. Since the hidden factors of observations are believed to have removed the intra-class difference introduced by the different modalities, we implement modality-independent face recognition based on the hidden factors. Given a probe observation $\vec{p}$ and a set of gallery observations $G = \{\vec{g}_1, , \vec{g}_m\}$, their L2-normalized inferred mutual components are $h_{\vec{p}}$ and $H_G = \{h_{\vec{g}_1}, \ldots, h_{\vec{g}_m}\}$, respectively. Denoting that $h_{\vec{g}_i} \sim N(\vec{\mu}_i, \Sigma_g)$ and $h_{\vec{p}} \sim N(\vec{\varphi}, \Sigma_p)$, the probability that $\vec{p}$ and $\vec{g}_i$ are associated with the same hidden factor (or mutual component) is:

$$P(\vec{\mu}_i) \propto \int \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_i)^T {\Sigma_g}^{-1} (\vec{x} - \vec{\mu}_i) \right) \exp\left(-\frac{1}{2} (\vec{x} - \vec{\varphi})^T {\Sigma_p}^{-1} (\vec{x} - \vec{\varphi}) \right) d\vec{x}.$$

We wish to select subject $\vec{g}_i \in G$ where $h_{\vec{g}_i} \sim N(\vec{\mu}_i, \Sigma_g)$, such that the $P(\vec{\mu}_i)$ can be maximized. According to the formula: $\int \exp\left(-\frac{1}{2} \vec{x}^T A \vec{x} + \vec{x}^T \vec{J} \right) d\vec{x} = \frac{(2\pi)^{n/2}}{\sqrt{|A|}} \exp\left(\frac{1}{2} \vec{J}^T A^{-1} \vec{J} \right)$, the expression can be simplified as:

$$\log(P(\vec{\mu}_i)) \propto w_i(\vec{\mu}_i) + \vec{\varphi}^T \vec{v}_i(\vec{\mu}_i), \tag{18}$$

where:

$$w_i(\vec{\mu}_i) = \frac{1}{2} \left[({\Sigma_g}^{-1} \vec{\mu}_i)^T ({\Sigma_g}^{-1} + {\Sigma_p}^{-1})^{-1} ({\Sigma_g}^{-1} \vec{\mu}_i) - \vec{\mu}_i^T {\Sigma_g}^{-1} \vec{\mu}_i \right],$$

$$\vec{v}_i(\vec{\mu}_i) = {\Sigma_p}^{-1}({\Sigma_g}^{-1} + {\Sigma_p}^{-1})^{-1}{\Sigma_g}^{-1}\vec{\mu}_i.$$

Although the expression has a complicated form, the computation can be implemented efficiently since both w_i and $\vec{v}_i$ are independent of $\vec{p}$, which means that they need to be computed only once for any $\vec{p}$ (although w_i and $\vec{v}_i$ depend on Σ_p, they are still independent of $\vec{p}$ since Σ_p depends only on the modality, not on the subject, according to (12)). We call it a linear classifier since it has a linear form when considering $\vec{\varphi}$ as the input.

3. INTEGRATED MULTI-CLASSIFIER FUSION FRAMEWORK

In practical applications, the test face images are usually subject to significant illumination and other changes, thus a single MCA-based classifier may lack the ability to model such complicated variations. To overcome this problem and improve recognition performance, we propose an MCA-based multi-classifier fusion framework. Figure 3 provides an overview of the MCA-based face recognition framework; we illustrate the framework with the sketch-photo matching problem, but it can also be applied to other cross-modality face recognition problems such as infrared-visible matching.

3.1. Local Feature Representations

Local facial features have been shown to be more effective than global facial features in representing face images at diverse scales and orientations [Mikolajczyk and Schmid 2005], hence we will use local feature representation in this study. Considering that the entire face image (which has high structural complexity) is difficult to characterize by a single image descriptor, we use a patch-based local feature representation scheme in this paper. We first divide the whole face image into a set of overlapping patches 16×16 pixels in size (overlapping factor = 0.5) and then apply local image descriptors to each patch. The extracted features from these patches are concatenated to form a long feature vector for further analysis. Of the existing local feature descriptors, Histograms of Oriented Gradients (HOG) [Dalal and Triggs 2005] and Multi-scale Local Binary Pattern (MLBP) [Mäenpää and Pietikäinen 2003] are among the most successful [Mikolajczyk and Schmid 2005]. We will therefore use both of them for feature description in our study. The HOG feature descriptor quantizes the gradient orientations into 12 directions with 30 degrees for each, and computes a histogram in which each bin corresponds to the gradient magnitude of that orientation. The accumulation of the histogram bins are weighted by the gradient. For the MLBP, we compute the MLBP descriptors with radii $\{1, 3, 5, 7\}$ using u2 coding [Ojala et al. 2002].

3.2. MCA-based Multi-classifier Fusion Framework

The resulting feature dimension is very high due to the use of overlapping patches and multi-scales in the representations. This makes the computational cost rather high. To overcome this problem, we construct multiple MCA models by slicing the long feature vector into a set of shorter local features and then integrating them to form a sophisticated system, as illustrated in Fig. 3. After slicing, we apply PCA on each slice to remove redundant information.

Suppose there are N slices in total with each one corresponding to one MCA-based classifier C_i. Assuming that these N classifiers are independent of each other, the combined posterior probability of these N classifiers is: $P = \prod_{i=1}^{N} P_i$, where P_i is the probability of the C_i (given the probe image and gallery image associated with the same hidden factor). According to (18), each classifier C_i outputs a score s_i that is proportional to $\log P_i$, and the combined score becomes:

$$s = \sum_{i=1}^{N} s_i. \tag{19}$$

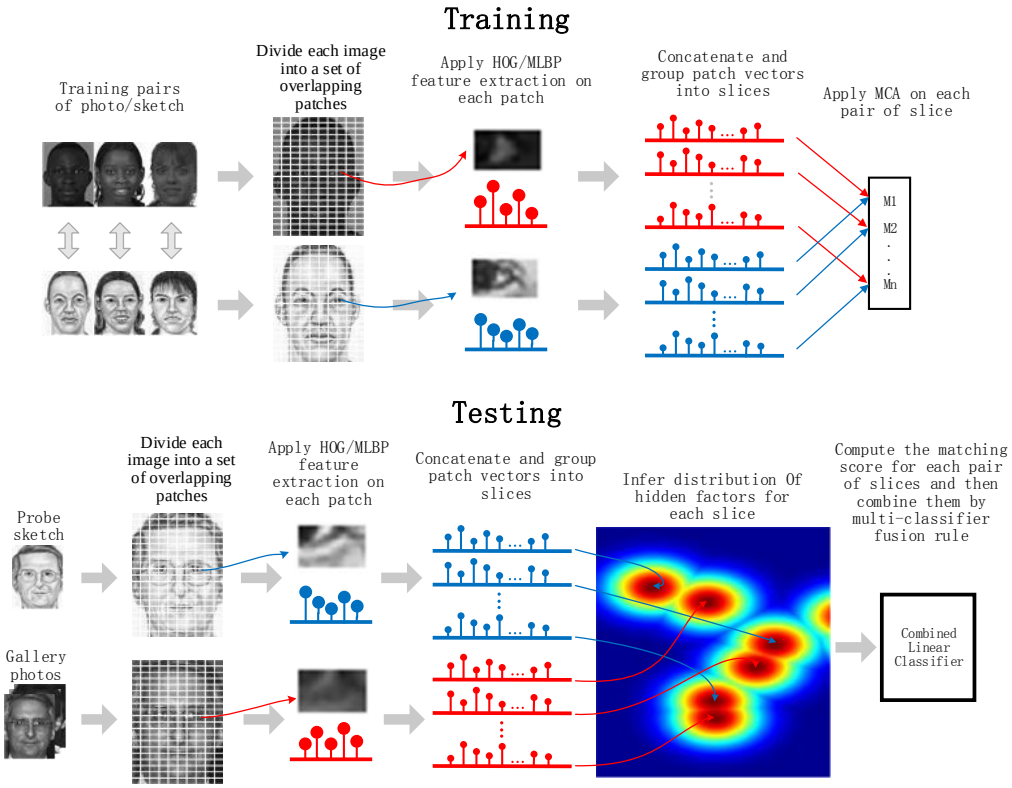

Fig. 3. Illustration for the pipeline of the MCA-based cross-modality face recognition system. Face features are extracted by applying HOG/MLBP descriptors on patches followed by grouping and slicing. In the training process, MCA algorithms are applied on each feature slice independently to obtain a series of MCA model M1, M2,..,Mn; In the testing process, we estimate the posterior distribution of hidden factor for each slice, and then match a probe face to the gallery faces with combined linear classifiers.

The procedure of our MCA-based multi-classifier fusion framework is summarized as follows.

In the training stage,

1. For each HOG or MLBP feature vector, divide the original vector equally into a set of feature slices without overlap.
2. Construct an MCA-based classifier based on each slice.

In the testing stage,

1. For each test sample which is represented by the two kinds of local features using the aforementioned local feature representation schemes, obtain the slices in the same way as in the training stage.
2. Use the trained MCA-based classifiers to determine the classification outputs of these slices.
3. Combine the outputs to obtain the final classification decision according to equation (19) for each test sample.

Discussion: The MCA-based multi-classifier fusion framework is an extension and enhancement of the MCA model. Compared to the original MCA approach, this fusion framework is more effective in handling the long feature vector. First, this fusion

framework is composed of several parallel classifiers with each one processing a portion of data, so efficient parallel computation is enabled. Second, the idea of separating the long feature vector into slices allows us to work on more manageable-sized data with respect to the number of training data. This will help to reduce the risk of overfitting and improve recognition performance. Our experimental results in Section IV also support the effectiveness of the fusion framework over the original MCA approach.

4. EXPERIMENTS

In this section, we conduct extensive experiments on two typical cross-modality face recognition scenarios: Sketch to Photo (SP) and Infrared to Visible (IV). The experimental protocol is as follows. Section 4.1 describes the experiments on CUHK Face-Sketch FERET database. We first compared our method against the state-of-the-art subspace methods on this database, as reported in Table I (rank-1 identification performance comparison) and Table II (verification performance comparison). Then we evaluated the performance of MCA-based multi-classifier fusion framework, as shown in Table III. Finally we compared our best result against the state-of-the-art results (provided by their papers) on this database. The same is true for Section 4.2 which describes the experiments on CUHK VIS-NIR face dataset. Section 4.3 describes the experiments on the enlarged gallery set (as reported in Figure 4), which is more close to the real-world face matching scenario. Section 4.4 describes the cross-database experiment validation. This is to demonstrate the generalization ability of our approach. Section 4.5 discusses the computational cost of our approach.

Dataset 1 - Sketch to Photo (SP). **The CUHK Face-Sketch FERET database (CUFSF)** [Zhang et al. 2011a] is used in this paper. It is the largest publicly available database for sketch-photo recognition and has some benchmark results. This database consists of sketches and photos of 1194 different persons, and each person has only one photo and one corresponding sketch composed by an artist. Sample photos and sketches are illustrated in Fig. 1(a), from which it can be seen that the sketches differ greatly from the corresponding photos. In addition, some textures of the sketches are lost and only a vague contour remains. In our experiments, 200 pairs of sketches and photos are selected as the training set to train our model, and the testing set is composed of the remaining 994 pairs.

Dataset 2 - Infrared to visible (IO). To systematically investigate this research topic in this study, we use a large visible and infrared database called **CUHK VIS-NIR face dataset**. This database consists of visible photos and infrared photos of 2876 different persons, with each person having one visible photo and one near infrared (NIR) photo. Samples of this database are shown in Fig.1 (b). We can see that although the near infrared photos maintain much more texture information compared to the sketches, the near infrared photos are blurred and have low contrast, and the gray distribution is also quite different from that of the visible photos. In our experiments, we divide the database into two parts: 1438 persons are used for training, and the remaining 1438 persons are used for testing. For the sketch-photo database, each image is cropped to size 200×160. For the infrared-visible database, each image is cropped to size 150×120. In addition, R^k at step (b) in the algorithm introduced in Section 3.2 is initialized as a random matrix with elements in [-0.1, 0.1]; and Σ^k is initialized as a diagonal matrix with diagonal elements equal to 0.01. To better evaluate the recognition performance with geometric and photometric interference filtered out, we preprocess the face images through the following steps: 1) rotate the face images to align them to the vertical face orientation; 2) scale the face images so that the distance between the two eyes is the same for all images; 3) crop the face images to remove the background and the hair region; 4) apply histogram equalization to the face images for photometric normalization.

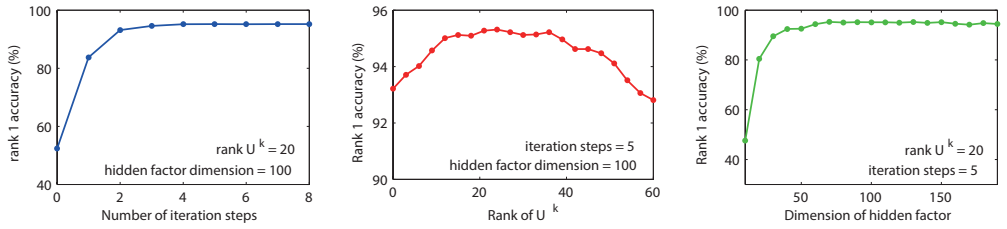

Fig. 4. Rank-1 recognition accuracy vs. (a) Number of iterations; (b) Rank of U^k; (c) Dimension of hidden factor. In (c), the dimension ranges from 10 to 190 with step=10. The HOG feature is used and the multi-classifier fusion method in section IV-B is applied with ♯ slices=50. The recognition rates are averaged over 10 random splits of 200 training persons and the remaining 994 testing persons on the SP database.

4.1. Experiment on CUHK Face-Sketch FERET database

4.1.1. Model parameter exploration. We first investigate the effect of various free parameters on the performance of the model, including the number of iteration steps for the EM algorithm, the rank of matrix U^k, and the dimension of hidden factors. Other factors are fixed when investigating one parameter. The HOG features with overlapping sampling factor of 0.5 and the window size of 16×16 pixels are used as face features in this experiment.

Number of iteration steps. The iteration steps can affect performance since the EM algorithm needs a number of steps to converge to a (locally) optimal solution. The number of iteration steps ranges from 0 to 8, and the rank 1 recognition accuracy is reported. As shown in Fig. 4(a), performance quickly improves as the number of iteration steps increases and converges to the optimal point within 5 steps, which indicates that our model can be trained efficiently.

Rank of U^k. The U^k explains the components for modality difference, and the rank r imposes a rank constraint on U^k. As we have noted, the model may unfortunately tend to over-fit the training data when the training set is not large enough. The introduction of r is intended to avoid the model becoming too complicated (in the field of machine learning, this means the VC dimension [Joachims 1998] of the system becomes too high) to adapt to many details of the training data. In this experiment, we explore rank r ranging from 0 to 60, and the rank-1 accuracy is reported in Fig. 4(b). There are three points we should note: 1) the modality term U^k has an active influence on recognition performance; 2) the model has a wide range for the rank of U^k to achieve good accuracy, which shows the robustness of our algorithm; 3) when the rank of U^k becomes too high, the model becomes over-fit (which is why we impose a rank constraint on U^k).

PCA + hidden factor dimension. We implement a PCA-based dimension reduction for the features prior to training. The PCA dimension is set as the number of training samples (in this case, the PCA dimension is 200 since we use 200 samples for training), which is enough to retain all the covariance of the training samples. The dimension of the hidden factors has an effect on performance since it represents the dimension of the hidden subspace from which we believe the observations are generated. We investigate the rank-1 recognition accuracy vs. hidden factor dimension, and the result is shown in Fig. 4(c). We find that the dimension of the hidden factors has a wide range from 70 to 150 to achieve good performance, which again reveals the robustness of our model.

4.1.2. Comparison Experiments. Since the proposed MCA approach can be considered as a novel subspace model, in this section we compare our mutual component analysis (MCA) approach with (i) several popular subspace analysis methods for regular face

Table I. Comparison of identification results (rank-1 identification rate) on the CUHK Face-Sketch FERET database.

Method	PCA	LDA	USA	PLS	CCA	LFDA	MCA
HOG	65.13%	86.72%	87.14%	87.92%	90.74%	90.49%	94.57%
MLBP	49.54%	74.02%	75.44%	77.38%	83.28%	82.27%	87.01%

Table II. Comparison of verification results on the CUHK Face-Sketch FERET database.

Method	PCA	LDA	USA	PLS	CCA	LFDA	MCA
HOG (VR0.1%FAR)	68.55%	87.94%	88.37%	90.32%	94.26%	92.87%	98.99%
MLBP (VR0.1%FAR)	58.75%	85.77%	86.41%	87.44%	90.87%	89.57%	94.67%
HOG (VR0.5%FAR)	76.37%	94.12%	95.31%	96.02%	98.89%	97.44%	100%
MLBP (VR0.5%FAR)	69.21%	90.53%	91.83%	93.11%	94.31%	95.32%	99.13%

recognition including principal component analysis (PCA) [Turk and Pentland 1991], linear discriminant analysis (LDA) [Belhumeur et al. 1997], unified subspace analysis (USA) [Wang and Tang 2004], and (ii) several newly-developed subspace methods in heterogeneous face recognition including PLS [Sharma and Jacobs 2011], CCA [Yi et al. 2007; Li et al. 2009], and LFDA [Klare et al. 2011], on two different local feature representations (HOG and MLBP, described in Section 3.1). The comparative results for face identification and face verification are reported in Table I and Table II respectively. From the results we make the following observations: 1) The relatively poor results of the traditional subspace methods on one hand show the difficulty of this database, on the other hand confirm the fact that traditional subspace methods are not suitable for the challenging cross-modality face recognition problem due to the great discrepancies between the two image styles. It is thus desirable to propose a new subspace model specifically tailored to this problem. 2) It is encouraging to see that our new approach, mutual component analysis (MCA), significantly outperforms all the other methods in different feature representations for both face identification and face verification by a large margin. Compared to several newly-developed subspace methods for heterogeneous face recognition (PLS [Sharma and Jacobs 2011], CCA [Yi et al. 2007; Li et al. 2009], and LFDA [Klare et al. 2011]) in particular, the proposed MCA approach obtains significant performance improvement. This shows the effectiveness of MCA.

The results of the MCA approach recorded in Table I and Table II are obtained by applying a single MCA-based classifier on the extracted local features (HOG or MLBP). As discussed in Section III, these results can be further improved by extending the MCA approach to a multi-classifier fusion framework. To verify this point and investigate the performance of our proposed MCA-based multi-classifier fusion framework, we conduct another comparative experiment. The results for both face identification and face verification are reported in Table III. From the results we have the following observations: 1) compared to the results in Table I and Table II, the developed MCA-based multi-classifier fusion framework boosts recognition performance, regardless of which local feature (MLBP or HOG) is used. This shows the advantage of the MCA-based multi-classifier fusion framework over the original MCA approach. 2) By combining the two kinds of local features, MLBP and HOG, we achieve the best results for both the identification and verification tasks. Considering the difficulty of this database, these are very encouraging results.

There are some state-of-the-art results available in the literature since the CUHK Face-Sketch FERET database [Zhang et al. 2011a] is a public domain database, hence

Table III. The results of the MCA-based multi-classifier fusion framework on the CUHK Face-Sketch FERET database.

Accuracy	MLBP (single classifier)	HOG (single classifier)	MLBP + HOG (multi-classifier)
Rank-1 identification rates	93.37%	98.00%	98.45%
Verification rates at FAR=0.1%	98.99%	99.52%	99.86%
Verification rates at FAR=0.5%	100%		

Table IV. Comparison of the state-of-the-art results with ours on the CUHK Face-Sketch FERET database. VR at 0.1% FAR are shown in this table.

Methods	Number of Subjects for Training	Number of Subjects for Testing	VR at 0.1%FAR
MRF + RS-LDA [Wang and Tang 2009]	500	694	29.54%
Kernel CSR (LBP) [Lei and Li 2009]			64.55%
Kernel CSR (SIFT) [Lei and Li 2009]			88.18%
LFDA [Klare et al. 2011]			90.78%
CITE [Zhang et al. 2011a]			98.70%
Regularized Discriminative Spectral Regression [Huang et al. 2013]			96.83%
Prototype random subspace [Klare and Jain 2013]			98.99%
OURS	200	994	99.86%

Table V. Comparison of identification results (rank-1 identification rate) on the CUHK VIS-NIR face dataset.

Method	PCA	LDA	USA	PLS	CCA	LFDA	MCA
HOG	31.22%	58.22%	59.04%	60.28%	64.83%	64.94%	70.12%
MLBP	38.57%	63.79%	65.23%	62.72%	68.70%	67.47%	76.42%

it is desirable to compare the state-of-the-art results on this database with ours. The comparative results are reported in Table IV. From these results, we can clearly see that our approach achieves superior performance over the state-of-the-art. Particularly noteworthy is the fact that our result is achieved on a more difficult testing scenario (using only 200 subjects as training and the other 994 subjects as testing) than the others (using 500 subjects as training and the remaining 694 subjects as testing). This further illustrates the effectiveness of MCA.

4.2. Experiment on CUHK VIS-NIR face dataset

The CUHK VIS-NIR face dataset contains 2876 different persons with each one having a pair of infrared and visible facial images. We use 1438 pairs of infrared and visible facial images as the training set and the remaining 1438 pairs as the testing set. We implement the comparative experiments similar to the CUHK Face-Sketch FERET database. The comparative results are reported in Table V-VIII. For the MCA-based multi-classifier fusion framework, the number of slices (for each HOG or MLBP feature vector) is set to 12. The results confirm our observations in the CUHK Face-Sketch FERET database. This further validates the effectiveness of the proposed approach.

4.3. Experiment on the Enlarged Gallery Set

To better investigate the performance of the proposed MCA model, we conduct an additional experiment on the enlarged gallery set by adding 10,000 images of 10,000 subjects from the MORPH database [Ricanek and Tesafaye 2006] to significantly increase the size of the gallery set, and compare our method with a top-performing heterogeneous face recognition approach in [Klare and Jain 2013]. Matching heterogeneous face images on a large-scale gallery set can produce results that more closely resemble a real-world face matching scenario. The identification rates are illustrated in Figure

Table VI. Comparison of verification results on the CUHK VIS-NIR face dataset.

Method	PCA	LDA	USA	PLS	CCA	LFDA	MCA
HOG (VR0.1%FAR)	37.24%	71.63%	74.42%	73.96%	81.21%	81.93%	83.37%
MLBP (VR0.1%FAR)	42.72%	81.22%	82.93%	82.07%	84.08%	83.49%	88.76%
HOG (VR0.5%FAR)	49.53%	81.42%	84.65%	82.46%	85.97%	86.23%	89.53%
MLBP (VR0.5%FAR)	54.85%	86.04%	87.55%	86.34%	90.14%	89.57%	93.21%

Table VII. The results of the MCA-based multi-classifier fusion framework on the CUHK VIS-NIR face dataset.

Accuracy	MLBP (single classifier)	HOG (single classifier)	MLBP + HOG (multi-classifier)
Rank-1 identification rates	84.03%	78.87%	86.43%
Verification rates at FAR=0.1%	91.42%	86.91%	93.12%
Verification rates at FAR=0.5%	95.85%	92.64%	96.82%

Table VIII. Comparison of our new approach with several newly-developed heterogeneous face recognition algorithms on the infrared-visible database. Listed are the Rank-1 identification accuracies.

Algorithm	Rank-1 Accuracy
LFDA [Klare et al. 2011]	69.22%
CITE [Zhang et al. 2011a]	72.53%
Regularized Discriminative Spectral Regression [Huang et al. 2013]	70.07%
Prototype random subspace [Klare and Jain 2013]	75.10%
Our approach	86.43%

5, from which we make the following observations. First, our approach consistently outperforms the method in [Klare and Jain 2013] by a clear margin. Second, the performance of our approach drops slightly more slowly than the method in [Klare and Jain 2013] as the gallery size enlarges, especially on the CUHK VIS-NIR database.

4.4. Cross-database Validation Experiment

Lastly, we conduct a cross-database experiment to verify the generalization ability of our approach and compare it with the approach in [Klare and Jain 2013], which is the best-performing heterogeneous face recognition approach in the literature. There are two public-domain face sketch databases: CUFS [Wang and Tang 2009] and CUFSF [Zhang et al. 2011a]. We use CUFS as the training database and CUFSF as the testing database. For the approach in [Klare and Jain 2013], we fuse three features: SIFT, MLBP and CSDN, as suggested by the paper. The comparative results for both identification and verification are reported in Table IX, and Figure 6. It is very clear that our approach significantly outperforms the approach (Prototype random subspace) in [Klare and Jain 2013] in the cross-database matching scenario for both the identification and verification tasks. This further confirms the effectiveness and generalization ability of our approach.

4.5. Computational Cost

In this subsection, we discuss the computational cost of our algorithm. The major computational cost of our algorithm comes from the computation of conditional moments given by (15) and (16). For the first conditional moment, the matrix multiplication dominates the cost, the complexity of which is bounded by $O(d^2N + dN^2 + N^3)$. For the

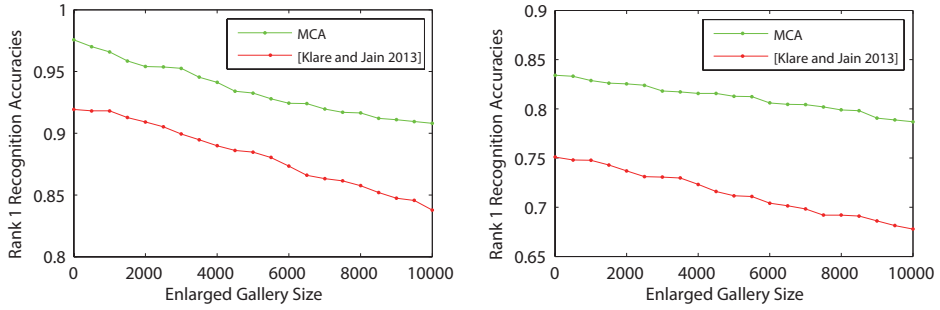

Fig. 5. The identification performance on the enlarged gallery set for (a) CUHK Face-Sketch FERET database and (b) CUHK VIS-NIR database. The size of the gallery set is increased by 500 at each step, up to 10000 images.

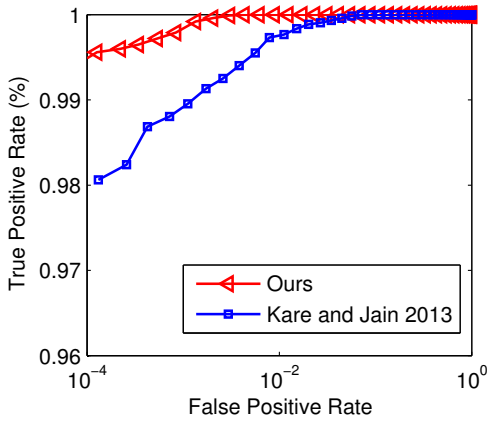

Fig. 6. The ROC comparison between Our approach and the top-performing approach (Prototype random subspace) in [Klare and Jain 2013] on the CUHK Face-Sketch FERET database.

Table IX. Comparative results for generalization ability performance.

Methods	Rank-1 Identification Rate	VR at 0.1%FAR	VR at 1%FAR
Prototype random subspace [Klare and Jain 2013]	62.40%	63.23%	82.16%
Our approach	69.10%	76.21%	92.71%

second conditional moment, it requires $O(d^3)$ for matrix inversion plus N outer product operations on d-dimensional vectors, which amounts to $O(d^3 + d^2N)$. Suppose the algorithm requires M iteration steps to come to a convergence, as $d = \alpha N$ where $\alpha < 1$ is a constant factor, then the complexity of the algorithm is $O(N^3M)$, where N is the number of training persons. In our experiments, the modeling process converges very quickly (within 8 iteration steps). The detailed convergence proof for our algorithm in the supplemental material provides more detail. Practically, on a 4-cores desktop PC running at 3.4GHz frequency, it takes around $9s$ to train a model for $1,400$ pairs of training images. Clearly, our approach can be efficiently used for practical applications.

5. CONCLUSIONS

In this paper, we have proposed a novel approach called mutual component analysis (MCA) for cross-modality face recognition. The key idea of the MCA approach is to learn a generative model to model the process of face image generation in different modalities. The learned generative model infers the hidden factors (invariant to modalities) that are only associated with a persons identity so that the cross-modality recognition problem can be well solved. To enhance the performance of MCA, we have further proposed an MCA-based multi-classifier fusion framework to boost recognition performance. Extensive experiments on two large cross-modality face databases show that our new approach obtains significant improvement over the state-of-the-art.

APPENDIX

Proof of Lemma 2.2:

Given model parameter estimates $\theta^K = (\vec{\beta}^K, T^K, \Sigma^K)$, drop the superscript and subscript of k and i for conciseness:

$$P_\theta(\vec{y}|\vec{m}) = P_\theta(\vec{m}|\vec{y})N(\vec{y}|\vec{0}, I).$$

Then:

$$\begin{aligned} P_\theta(\vec{y}|\vec{m}) &\propto \exp\left(\vec{y}^T T^T \Sigma^{-1}(\vec{m} - \vec{\beta}) - \frac{\vec{y}^T T^T \Sigma^{-1} T\vec{y}}{2} - \frac{\vec{y}^T \vec{y}}{2}\right) \\ &\propto \exp\left(\vec{y}^T T^T \Sigma^{-1}(\vec{m} - \vec{\beta}) - \frac{\vec{y}^T L\vec{y}}{2}\right) \\ &\propto \exp\left(-\frac{1}{2}(\vec{y} - \vec{\varphi})^T L(\vec{y} - \vec{\varphi})\right). \end{aligned}$$

□

Proof of Proposition 2.1:

According to equation (9), we have:

$$\vec{v}_i = S\vec{y}_i, \tag{20}$$

where $\vec{v}_i = \frac{1}{K}\sum_{k=1}^{K}(\vec{m}_i^k - \vec{\beta}^k)$. According to (20), multiplying their transposes on both sides accordingly, and summing up them gives:

$$\frac{1}{N}\sum_{i=1}^{N}\vec{v}_i\vec{v}_i^T = S\left(\frac{1}{N}\sum_{i=1}^{N}\vec{y}_i\vec{y}_i^T\right)S^T. \tag{21}$$

Since $\vec{y}_i^T \sim N(0, I)$, $\frac{1}{N}\sum_{i=1}^{N}\vec{y}_i\vec{y}_i^T$ becomes an identity matrix. The left part of (21) corresponds to the covariance matrix of the observations: $CM = \frac{1}{N}\sum_{i=1}^{N}\vec{v}_i\vec{v}_i^T$. Accordingly (21) can be re-written as:

$$CM = SS^T. \tag{22}$$

If $\vec{e}$ is an eigenvector of the symmetric covariance matrix CM, then:

$$CM = \vec{e}D\vec{e}^T, \tag{23}$$

where D is a diagonal matrix whose elements consist of eigenvalues of CM. Combining (22) and (23), we can see that $S = V\sqrt{D}$, where V is a $D \times d$ matrix whose columns consist of the eigenvectors of CM corresponding to the largest d eigenvalues. □

Proof of Proposition 2.3:
Let $\theta_0^k = (R_0^k, \Sigma_0^k)$ be the initial estimates, and $\theta^k = (R^k, \Sigma^k)$ be the new estimates. We wish to select new model parameters θ^k, such that the joint probability can be maximized:

$$\theta^k = \arg\max_{\theta^k} \prod_{k=1}^{K} P_{\theta^k}(M^{1,\ldots,K}, \vec{Y}). \tag{24}$$

Since the observations of different modalities are independent, we have:

$$\theta^k = \arg\max_{\theta^k} \prod_{k=1}^{K} P_{\theta^k}(M^k, \vec{Y}).$$

Since the observations of different subjects are also independent, we further have:

$$P_{\theta^k}(M^k|\vec{Y}) = \prod_{i=1}^{N}\prod_{k=1}^{K} P_{\theta^k}(\vec{m}_i^k|\vec{y}_i).$$

Since $P(\vec{y}_i) \sim N(\vec{0}|I)$, thus:

$$P_{\theta^k}(M^k|\vec{Y}) = \prod_{i=1}^{N}\prod_{k=1}^{K} P_{\theta^k}(\vec{m}_i^k|\vec{y}_i)N(\vec{0}|I). \tag{25}$$

Maximizing (24) is equivalent to maximizing:

$$\theta^k = \arg\max_{\theta^k} \left\{\log P_{\theta^k}(M^k, \vec{Y})\right\}. \tag{26}$$

Substituting (25) into (26) results in:

$$\theta^k = \arg\max_{\theta^k} \sum_{k=1}^{K}\sum_{i=1}^{N} \log P_{\theta^k}(\vec{m}_i^k|\vec{y}_i). \tag{27}$$

Since $\vec{y}_i$ can only be inferred by the initial model parameters and observations according to Lemma 2.2, the most natural way to estimate θ^k is to maximize the expectation of $\log P_{\theta^k}(\vec{m}_i^k|\vec{y}_i)$:

$$\theta^k = \arg\max_{\theta^k} \sum_{k=1}^{K}\sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \log P_{\theta^k}(\vec{m}_i^k|\vec{y}_i) d\vec{y}_i.$$

Define a function of the model parameter θ as follows:

$$F(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \log P_{\theta^k}(\vec{m}_i^k|\vec{y}_i) d\vec{y}_i.$$

Thus, our aim is to select θ^* as the new estimates so as to maximize $F(\theta)$.
a) Optimization of R^k
Since

$$\frac{\partial F(\theta)}{\partial R^k} = \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \frac{\partial \log P_{\theta^k}(\vec{m}_i^k|\vec{y}_i)}{\partial R^k} d\vec{y}_i,$$

and

$$P_{\theta^k}(\vec{m}_i^k|\vec{y}_i) \sim N(\vec{m}_i^k|\vec{\beta}^k + S\vec{y}_i + R^k W\vec{y}_i, \Sigma^k),$$

we have

$$\frac{\partial F(\theta)}{\partial R^k} = \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \frac{\partial}{\partial R^k} \left\{ -\frac{1}{2}(D\log(2\pi) + \log|\Sigma^k| + \vec{v}^{k\,T} \Sigma^{k^{-1}} \vec{v}^k) \right\} d\vec{y}_i, \quad (28)$$

where $\vec{v}^k = \vec{m}_i^k - \vec{\beta}^k - S\vec{y}_i - R^k W \vec{y}_i$.
Since

$$\frac{\partial}{\partial R^k(m,n)} \left\{ -\frac{1}{2}(D\log(2\pi) + \log|\Sigma^k| + \vec{v}^{k\,T} \Sigma^{k^{-1}} \vec{v}^k) \right\} = \sum_{s=1}^{D} \frac{\partial(\vec{v}^{k\,T} \Sigma^{k^{-1}} \vec{v}^k)}{\partial \vec{v}^k(s)} \frac{\partial \vec{v}^k(s)}{\partial R^k(m,n)},$$

according to the general formula of vector derivatives:

$$\frac{\partial(\vec{x}^T P \vec{x})}{\vec{x}} = 2P\vec{x},$$

we have:

$$\frac{\partial(\vec{v}^{k\,T} \Sigma^{k^{-1}} \vec{v}^k)}{\vec{v}^k} = 2\Sigma^{k^{-1}} \vec{v}^k.$$

Also:

$$\frac{\partial \vec{v}^k(i)}{\partial R^k(m,n)} = \begin{cases} \sum_{j=1}^{d} W(n,j) * \vec{y}_i(j), & i = m; \\ 0, & i \neq m. \end{cases}$$

Then substituting the above into (28) gives:

$$\frac{\partial F(\theta)}{\partial R^k} = \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \Sigma^{k^{-1}} \vec{v}^k \vec{y}_i^T W^T d\vec{y}_i. \quad (29)$$

Let $\frac{\partial F(\theta)}{\partial R^k} = 0$,

$$\Sigma^{k^{-1}} \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \left[(\vec{m}_i^k - \vec{\beta}^k)\vec{y}_i^T - (S + R^k W)\vec{y}_i \vec{y}_i^T \right] d\vec{y}_i = 0.$$

Thus:

$$\sum_{i=1}^{N} (\vec{m}_i^k - \vec{\beta}^k)\ \vec{E}_{M_i}^{1\,T} = (S + R^k W) \sum_{i=1}^{N} \vec{E}_{M_i}^2$$

where

$$\vec{E}_{M_i}^1 = \int P_{\theta_0^k}(\vec{y}_i|M_i)\vec{y}_i d\vec{y}_i,$$
$$\vec{E}_{M_i}^2 = \int P_{\theta_0^k}(\vec{y}_i|M_i)\vec{y}_i \vec{y}_i^T d\vec{y}_i.$$

Since:

$$P_{\theta_0^k}(\vec{y}_i|M_i) = \frac{1}{K} \sum_{k=1}^{K} P_{\theta_0^k}(\vec{y}_i|\vec{m}_i^k),$$

we have:

$$\vec{E}_{M_i}^1 = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \vec{E}_{M_i}^{1,k},$$
$$\vec{E}_{M_i}^2 = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \vec{E}_{M_i}^{2,k}.$$

According to Lemma 2.2:

$$\vec{E}_{M_i}^{1,k} = {L^k}^{-1}(S + R_0^k W)^T \ {\Sigma_0^k}^{-1}(\vec{m}_i^k - \vec{\beta}^k),$$
$$\vec{E}_{M_i}^{2,k} = {L^k}^{-1} + \vec{E}_{M_i}^{1,k} \ {\vec{E}_{M_i}^{1,k}}^T.$$

which is required by (13).

b) Optimization of Σ^k

$$\frac{\partial F(\theta)}{\partial \Sigma^k} = \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \frac{\partial}{\partial \Sigma^k} \sum_k (\log|\Sigma^k| + \ {\vec{v}^k}^T {\Sigma^k}^{-1} \vec{v}^k) d\vec{y}_i. \tag{30}$$

According to the general formula of matrix derivatives:

$$\tfrac{\partial}{\partial A} \log|A| = (A^T)^{-1},$$
$$\tfrac{\partial}{\partial A} trace(BA^{-1}C) = -(A^{-1}CBA^{-1})^T.$$

we have:

$$\tfrac{\partial}{\partial \Sigma^k} \log|\Sigma^k| = ({\Sigma^k}^T)^{-1},$$
$$\tfrac{\partial}{\partial \Sigma^k} ({\vec{v}^k}^T {\Sigma^k}^{-1} \vec{v}^k) = -({\Sigma^k}^{-1} \vec{v}^k \ {\vec{v}^k}^T {\Sigma^k}^{-1})^T.$$

Substituting them into (30) results in:

$$\frac{\partial F(\theta)}{\partial \Sigma^k} = \sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \left(({\Sigma^k}^T)^{-1} - ({\Sigma^k}^{-1} \vec{v}^k \ {\vec{v}^k}^T {\Sigma^k}^{-1})^T \right) d\vec{y}_i.$$

Enforcing $\frac{\partial F(\theta)}{\partial \Sigma^k} = 0$ and considering Σ^k is symmetric, it follows:

$$N{\Sigma^k}^{-1} - {\Sigma^k}^{-1} \left(\sum_{i=1}^{N} \int P_{\theta_0^k}(\vec{y}_i|M_i) \left(\vec{v}^k \ {\vec{v}^k}^T \right) d\vec{y}_i \right) {\Sigma^k}^{-1} = 0.$$

According to (13):

$$\sum_{i=1}^{N} (\vec{m}_i^k - \vec{\beta}^k) \ {\vec{E}_{M_i}^1}^T = T^k \sum_{i=1}^{N} \vec{E}_{M_i}^2$$

we have

$$\Sigma^k = \frac{1}{N} \sum_{i=1}^{N} \left((\vec{m}_i^k - \vec{\beta}^k)(\vec{m}_i^k - \vec{\beta}^k)^T - T^k \vec{E}_{M_i}^1 (\vec{m}_i^k - \vec{\beta}^k)^T \right).$$

which is required by (14). □

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

REFERENCES

Peter N. Belhumeur, João P Hespanha, and David Kriegman. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7 (1997), 711–720.

Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. 2012. Distinguishing Facial Features for Ethnicity-Based 3D Face Recognition. *ACM Trans. Intell. Syst. Technol.* 3, 3, Article 45 (May 2012), 20 pages. DOI:http://dx.doi.org/10.1145/2168752.2168759

Rama Chellappa, Charles L Wilson, and Saad Sirohey. 1995. Human and machine recognition of faces: A survey. *Proc. of the IEEE* 83, 5 (1995), 705–741.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 1. IEEE, 886–893.

Changxing Ding, Chang Xu, and Dacheng Tao. 2015. Multi-task pose-invariant face recognition. *IEEE Trans. Image Process.* 24, 3 (2015), 980–993.

Zhengming Ding, Shao Ming, and Yun Fu. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI Conference on Artificial Intelligence*. AAAI.

Ralph Ewerth, Markus Mühling, and Bernd Freisleben. 2012. Robust Video Content Analysis via Transductive Learning. *ACM Trans. Intell. Syst. Technol.* 3, 3, Article 41 (May 2012), 26 pages. DOI:http://dx.doi.org/10.1145/2168752.2168755

Cong Geng and Xudong Jiang. 2011. Face recognition based on the multi-scale local image structures. *Pattern Recognition* 44, 10 (2011), 2565–2575.

Xiangsheng Huang, Zhen Lei, Mingyu Fan, Xiao Wang, and Stan Z Li. 2013. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Trans. Image Processing* 22, 1 (2013), 353–362.

Kui Jia and Shaogang Gong. 2008. Generalized Face Super-Resolution. *IEEE Trans. Image Process.* (2008).

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proc. Eur. Conf. Mach. Learn.*, Claire Ndellec and Cline Rouveirol (Eds.). Lecture Notes in Computer Science, Vol. 1398. Springer Berlin Heidelberg, 137–142. DOI:http://dx.doi.org/10.1007/BFb0026683

Brendan Klare and Anil K Jain. 2010. Heterogeneous face recognition: Matching NIR to visible light images. In *Proc. Int. Conf. Pattern Recognit.* IEEE, 1513–1516.

Brendan F Klare and Anil K Jain. 2013. Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 6 (2013), 1410–1422.

Brendan F Klare, Zhifeng Li, and Anil K Jain. 2011. Matching forensic sketches to mug shot photos. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 3 (2011), 639–646.

Zhen Lei and Stan Z Li. 2009. Coupled spectral regression for matching heterogeneous faces. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 1123–1128.

Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. 2009. Maximizing intra-individual correlations for face recognition across pose differences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 605–611.

Stan Z Li. 2009. Heterogeneous face biometrics. In *Encyclopedia of Biometrics*. Springer, 700–702.

Yuelong Li, Li Meng, Jufu Feng, and Jigang Wu. 2014. Downsampling sparse representation and discriminant information aided occluded face recognition. *SCIENCE CHINA Information Sciences.* 57 (2014).

Zhifeng Li, Dihong Gong, Xuelong Li, and Dacheng Tao. 2015. Learning Compact Feature Descriptor and Adaptive Matching Framework for Face Recognition. *IEEE Trans. Image Process.* 24, 9 (2015), 2736–2745.

Zhifeng Li, Dihong Gong, Yu Qiao, and Dacheng Tao. 2014. Common feature discriminant analysis for matching infrared face images to optical face images. *IEEE Trans. Image Process.* 23, 6 (2014), 2436–2445.

Zhifeng Li, Dahua Lin, and Xiaoou Tang. 2009. Nonparametric Discriminant Analysis for Face Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence.* 31, 4 (2009), 755–761.

Zhifeng Li, Unsang Park, and Anil. K. Jain. 2011. A Discriminative Model for Age Invariant Face Recognition. *IEEE Trans. Information Forensics and Security.* 6, 3 (2011), 1028–1037.

Zhifeng Li and Xiaoou Tang. 2007. Using Support Vector Machines to Enhance the Performance of Bayesian Face Recognition. *IEEE Trans. Information Forensics and Security.* 2, 2 (2007), 174–180.

Shengcai Liao, Dong Yi, Zhen Lei, Rui Qin, and Stan Z Li. 2009. Heterogeneous face recognition from local structures of normalized appearance. In *Advances in Biometrics*. Springer, 209–218.

Dahua Lin and Xiaoou Tang. 2006. Inter-modality face recognition. In *Proc. Eur. Conf. Comput. Vis.* Springer, 13–26.

Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. 2005. A nonlinear approach for face sketch synthesis and recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 1. IEEE, 1005–1010.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision* 60, 2 (2004), 91–110.

Topi Mäenpää and Matti Pietikäinen. 2003. Multi-scale binary patterns for texture analysis. In *Image Analysis*. Springer, 885–892.

Krystian Mikolajczyk and Cordelia Schmid. 2005. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 10 (2005), 1615–1630.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (2002), 971–987.

Karl Ricanek and Tamirat Tesafaye. 2006. Morph: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*. IEEE, 341–345.

Ling Shao, Li Liu, and Xuelong Li. 2014a. Feature learning for image classification via multiobjective genetic programming. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 7 (2014), 1359–1371.

Ling Shao, Di Wu, and Xuelong Li. 2014b. Learning deep and wide: A spectral method for learning deep networks. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 12 (2014), 2303–2308.

Ming Shao, Zhengming Ding, and Yun Fu. 2015. Sparse Low-Rank Fusion based Deep Features for Missing Modality Face Recognition. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. IEEE.

Ming Shao and Yun Fu. 2013. Hierarchical hyperlingual-words for multi-modality face classification. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (2013).

Abhishek Sharma and David W Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 593–600.

Xiaoou Tang and Zhifeng Li. 2009. Audio-guided Video Based Face Recognition. *IEEE Trans. Circuits and Systems for Video Technology.* 19, 7 (2009), 955–964.

Xiaoou Tang and Xiaogang Wang. 2004. Face sketch recognition. *IEEE Trans. Circuits Syst. Video Technol.* 14, 1 (2004), 50–57.

Joshua B Tenenbaum and William T Freeman. 2000. Separating style and content with bilinear models. *Neural Computation* 12, 6 (2000), 1247–1283.

Matthew A Turk and Alex P Pentland. 1991. Face recognition using eigenfaces. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 586–591.

Xiaogang Wang and Xiaoou Tang. 2004. A unified framework for subspace face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 9 (2004), 1222–1228.

Xiaogang Wang and Xiaoou Tang. 2009. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 11 (2009), 1955–1967.

Lior Wolf, Tal Hassner, Yaniv Taigman, and others. 2008. Descriptor based methods in the wild. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

John Wright and Gang Hua. 2009. Implicit elastic matching with random projections for pose-variant face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 1502–1509.

Ruomei Yan, Ling Shao, and Yan Liu. 2013. Nonlocal hierarchical dictionary learning using wavelets for image denoising. *IEEE Trans. Image Process.* 22, 12 (2013), 4689–4698.

Dong Yi, Rong Liu, RuFeng Chu, Zhen Lei, and Stan Z Li. 2007. Face matching between near infrared and visible light images. In *Advances in Biometrics*. Springer, 523–530.

Lei Zhang, Xiantong Zhen, and Ling Shao. 2014. Learning object-to-class kernels for scene classification. *IEEE Trans. Image Process.* 23, 8 (2014), 3241–3253.

Wei Zhang, Xiaogang Wang, and Xiaoou Tang. 2011a. Coupled information-theoretic encoding for face photo-sketch recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 513–520.

Zeda Zhang, Yunhong Wang, and Zhaoxiang Zhang. 2011b. Face synthesis from near-infrared to visual light via sparse representation. *International Joint Conference on Biometrics* (2011).

W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. 2003. Face Recognition: A Literature Survey. *ACM Comput. Surv.* 35, 4 (Dec. 2003), 399–458. DOI:http://dx.doi.org/10.1145/954339.954342

Xiantong Zhen, Ling Shao, and Xuelong Li. 2014. Action recognition by spatio-temporal oriented energies. *Information Sciences* 281 (2014), 295–309.

Received February XXXX; revised March XXXX; accepted June XXXX