Matching NIR Face to VIS Face Using Transduction

Jun-Yong Zhu, *Student Member, IEEE*, Wei-Shi Zheng, *Member, IEEE*,
Jian-Huang Lai, *Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*— Visual versus near infrared (VIS-NIR) face image matching uses an NIR face image as the probe and conventional VIS face images as enrollment. It takes advantage of the NIR face technology in tackling illumination changes and low-light condition and can cater for more applications where the enrollment is done using VIS face images such as ID card photos. Existing VIS-NIR techniques assume that during classifier learning, the VIS images of each target people have their NIR counterparts. However, since corresponding VIS-NIR image pairs of the same people are not always available, which is often the case, so those methods cannot be applied. To address this problem, we propose a transductive method named transductive heterogeneous face matching (THFM) to adapt the VIS-NIR matching learned from training with available image pairs to all people in the target set. In addition, we propose a simple feature representation for effective VIS-NIR matching, which can be computed in three steps, namely Log-DoG filtering, local encoding, and uniform feature normalization, to reduce heterogeneities between VIS and NIR images. The transduction approach can reduce the domain difference due to heterogeneous data and learn the discriminative model for target people simultaneously. To the best of our knowledge, it is the first attempt to formulate the VIS-NIR matching using transduction to address the generalization problem for matching. Experimental results validate the effectiveness of our proposed method on the heterogeneous face biometric databases.**

Manuscript received November 2, 2013; accepted December 28, 2013. Date of publication January 13, 2014; date of current version February 14, 2014. This work was supported in part by Guangdong Provincial Government of China through the Computational Science Innovative Research Team Program, in part by NSFC under Grants 61102111 and 61173084, in part by Guangdong Natural Science Foundation under Grant S2012010009926, in part by the 12th Five-year Plan China S&T Supporting Programme under Grant 2012BAK16B06, in part by the Specialized Research Fund for the Doctoral Program of Higher Education 20110171120051, in part by the Fundamental Research Funds for the Central Universities under Grant 12lgpy28, in part by Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, and in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sviatoslav S. Voloshynovskiy. *(Corresponding author: W.-S. Zheng.)*

J.-Y. Zhu is with the School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, China, and also with SYSU-CMU Shunde International Joint Research Institute, Shunde, China (e-mail: jonesjunyong@gmail.com).

W.-S. Zheng is with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China, and also with the Guangdong Province Key Laboratory of Computational Science, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

J.-H. Lai is with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China, the Guangdong Province Key Laboratory of Computational Science, Guangzhou 510275, China, and also with SYSU-CMU Shunde International Joint Research Institute, Shunde, China (e-mail: stsljh@mail.sysu.edu.cn).

S. Z. Li is with the Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China (e-mail: szlig@cbsr.ia.ac.cn).

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

Digital Object Identifier 10.1109/TIFS.2014.2299977

***Index Terms*— Heterogeneous face recognition, VIS-NIR face matching, transductive learning.**

I. Introduction

FACE Recognition has been active for the last two decades due to its wide range of applications in law enforcement and verification systems. However, previous works primarily focus on visible (VIS) images, whose imaging is within the wavelength range from $0.4\mu m$ to $0.7\mu m$. Despite great success achieved in controlled environment, face recognition is still a challenging problem under uncontrolled illumination condition. Recently, developing face recognition system based on active near infrared (NIR) images, whose imaging is from $0.7\mu m$ to $1.1\mu m$, is proved to be less sensitive to visible light illumination changes [17]. As a result, identification of individuals via NIR imaging technique can significantly enhance the performance of recognition system in low light conditions. Such characteristic can be critical and demanded for some applications, especially for security surveillance in low light conditions [7]. For instance, the entry verification at arrival hall, identity verification at a ATM machine, E-passport, machine readable traveling document (MRTD) and etc [13], [14], [18], [30], [31].

However, there is an obstacle for comparing the captured NIR images with existing face images in our security sectors, as many largely deployed face recognition systems almost exclusively require individuals to get enrolled using their VIS face images. Hence, it is important to develop technologies for addressing the matching problem between NIR and VIS face images, and this is called the *VIS-NIR face matching problem* [17]. The latest multi-biometric grand challenge (MBGC 2008) has also set up a new test to investigate the matching between VIS images and the partially occluded faces in the NIR videos. By matching NIR probe images to the gallery VIS images, we can extend the VIS-VIS face recognition system to a heterogeneous NIR-VIS faces matching system in order to take advantages of NIR images to help face recognition in poor illumination conditions.

How to alleviate the gap between VIS and NIR domains is the main challenge of VIS-NIR face matching. Most existing methods try to achieve this goal via subspace learning methods [14], [19], [31] or by designing invariant descriptors via local feature approaches [13], [18]. These works assume that target people are included in the training set and for each individual there are labeled VIS-NIR pairs for learning the matching, which is an inductive learning procedure. However, since most target people have got enrolled only using VIS

images in existing security systems, i.e. the target VIS images do not have the corresponding NIR mates labeled and registered in the system, the matching learned for training may not be optimal for the target people, and this domain gap has not been considered in the existing inductive VIS-NIR matching approaches. In order to solve this problem, we reformulate the heterogeneous face matching problem using transduction in order to transfer the matching model learned from training to the target people.

To be detailed, we mainly address the VIS-NIR matching problem by introducing a transductive model, namely, *Transductive Heterogeneous Face Matching (THFM)*, which is shown in Fig. 1. We adapt the VIS-NIR matching learned from training to target, so that the VIS-NIR matching for target people can be considered to alleviate the domain variation. In other words, we assume a set of VIS-NIR pairs of training people are available and guide the learned VIS-NIR matching upon training to facilitate the matching for target ones. Our domain adaptation differs from existing ones [3], [22] mainly in two-fold: 1) our domain adaptation is related to two modalities rather than one, and we aim to adapt the probe NIR faces to the gallery VIS faces for the target people only; and 2) there are heterogeneous difference between VIS and NIR images, leading to the multi-modality distribution for each individual. In order to tackle the heterogeneous difference, we further contribute to proposing a novel descriptor by involving Log-DoG filtering and local encoding (e.g. LBP, HOG), which is insensitive to light source in order to assist our matching adaptation. More importantly, we provide theoretical analysis based on the illumination reflectance model.

As far as we know, it is the first attempt to develop a transductive VIS-NIR matching model. Although there is an existing work on transductive face recognition [15], the motivations are different, and more importantly they cannot address the heterogeneous face matching problem. In summary, the novelties of this paper are in four-folds.

1) We formulate the VIS-NIR face matching problem in a transductive framework, which is useful to reduce the modality gap for the target data.
2) We provide a comprehensive analysis on the feature representation and prove that LBP and HOG codings on both original images and Log-DoG filtered images are somehow invariant (insensitive) to the source light change under certain assumption, i.e. relatively consistent across VIS and NIR domains.
3) By simultaneously extracting domain invariant and target-related discriminative features, we propose the THFM approach using transduction. We also validate that our transductive method can work under the out-of-sample setting.
4) Impressive results (**VR** 98.42% at **FAR** 0.1%) are grained on VIS-NIR face database using our proposed feature representation and the learning framework THFM.

II. Related Work

Matching VIS-NIR images is challenging mainly due to the matching across different image modalities. We review related works in three aspects: (i) face synthesis analysis; (ii) subspace methods; (iii) local feature based approaches.

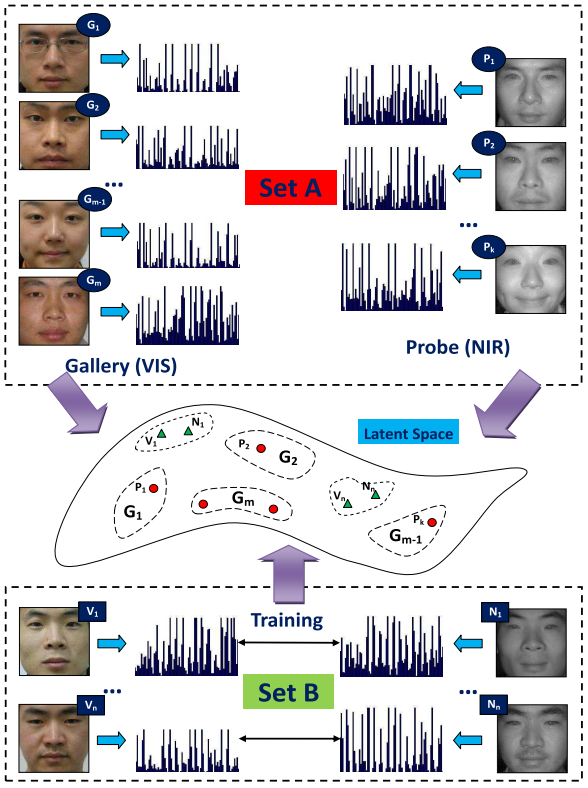

Fig. 1. Transductive VIS-NIR Face Matching Framework. We consider generalization of VIS face recognition system to a VIS-NIR matching system using a handful of training data. Target people in the Gallery are denoted as Set A and training people are denoted as Set B. The (VIS) faces in top left belong to the original VIS system, and probe NIR faces in top right are captured by online multi-cameras. Our goal is to search for latent space in which the heterogeneous face matching problem can be casted to the classical homogeneous face recognition.

Face synthesis analysis. Wang et al. in [26] transformed face image from one type to another using *face analogy* and subsequently matched synthetic query images to gallery set. Following the idea that a patch of image has nearly the same similarity as its neighboring patches in VIS and the corresponding NIR domains, Zhang et al. [33] developed a face synthesis approach via sparse representation, where a pair of over-complete dictionary are learned and the corresponding sparse coefficients of VIS and NIR images are assumed to be similar. However, synthetic methods require lots of well registered pairwise VIS-NIR face images for learning the mapping, which is hard to meet when only limited labeled VIS-NIR images are available.

Subspace Methods. Different from peer to peer transformation, subspace methods focus on searching a common subspace where VIS and NIR images from the same individual have similar representations. Yi et al. introduced canonical correlation analysis (CCA) to learn the correlation between NIR and VIS faces from NIR-VIS face pairs [31]. In order to tackle the inter-modality problem, Lin and Tang [19] considered the empirical discriminative power and the local smoothness of feature transformation and proposed a common discriminant feature extraction (CDFE), in which both

inter-modality discriminant information and intra-modality local smoothness were involved. Recently, Lei and Li [14] suggested solving the problem via coupled spectral regression (CSR). In their model, a low dimensional representation for each face was first computed using discriminative graph embedding method and then two associated projections were learned respectively to project heterogeneous data into the discriminative common subspace for final classification. Our work also mines a subspace, but our objective is for modeling domain adaptation for VIS-NIR matching in a transductive way, while these related works are non-transductive.

Invariant Feature Extraction. The local feature based methods aim at exploring invariant features to lighting conditions. Inspired by Tan and Triggs [24], Goswami et al. introduced an effective preprocessing chain to reduce the difference between VIS ans NIR facial images based on Gamma correction, Difference-of-Gaussian (DoG) filtering and contrast equalization [11]. Liao et al. suggested encoding both VIS and NIR face images using Multi-block LBP (MB-LBP) followed by DoG filtering [18]. Gentle AdaBoost and R-LDA were conducted for further feature selection. Following this work, Klare et al. incorporated histograms of oriented gradients (HOG) feature descriptor combined with LBP descriptor to describe each face and obtained significant improvement [13]. Binary Laplacian of Gaussian (LoG) was also investigated in [30]. Recently, Liu et al. proposed Light Source Invariant Features (LSIFs) to fill the gap between VIS and NIR face images [20]. In this work, multi-scale DoG is first performed to generate over-complete face representation, and then three local descriptors namely HOG, GLOH and SIFT are applied to construct the candidate feature pool, and finally Gentle AdaBoost is used to select the best features. However, AdaBoost is time consuming and needs abundant samples for obtaining robust performance, and it limits its use in our case. Alternatively, in order to assist the learning model in this work, we are more willing to exploit the domain invariant feature in a more efficient way using learning procedure like [13]. More importantly, existing feature descriptors are designed empirically and lacking of theoretical support. In this work, we explore the fundamental of the rationale of some popular existing descriptors for VIS-NIR matching and further introduce our proposed descriptor along with illumination invariant property analysis.

III. Preliminaries

A. Problem Formulation

We consider generalization of VIS face system to a VIS-NIR matching system, which can make existing VIS face recognition systems be able to cope with the newly input NIR images. As aforementioned, to realize the generalization, we attempt to adapt the matching learned from training (with pairwise VIS and NIR images) to target people (with labeled VIS gallery images and unlabeled NIR images to be matched).

More specifically, we have a set of training data consisting of pairwise VIS-NIR face images denoted as $\{x_{q,i}^{S\_VIS}|q \in C_S\}$ and $\{x_{q,j}^{S\_NIR}|q \in C_S\}$ respectively, where $x_{q,i}^{S\_VIS}$ is the i^{th} VIS image of class q in the training set, and C_S represents the set containing all training classes. For target people, there are a gallery set of labeled VIS images and a probe set of unlabeled NIR image. The gallery set is denoted by $\{x_{p,i}^{Gallery}|p \in C_T\}$ enrolled in an existing face recognition system, where $x_{p,i}^{Gallery}$ is the i^{th} sample of class p in the gallery set, C_T denotes the set of target people (denoted as Set A in Fig. 1). Our task is to match any NIR image $x_{p,j}^{Probe}$ in the probe set to its corresponding gallery image $x_{p,i}^{Gallery}$ in VIS domain, where $x_{p,j}^{Probe}$ and $x_{p,i}^{Gallery}$ belong to the same person. We realize this generalization by learning a transductive framework. The whole VIS-NIR face matching framework is illustrated in Fig. 1. Note that the labels of gallery samples are available but those of probe set are unknown.

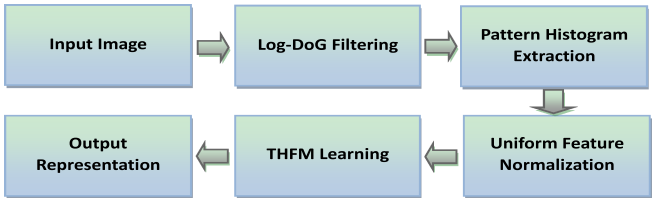

Fig. 2. Flowchart of the proposed transductive heterogeneous face matching framework.

B. Face Recognition Using Transduction

Different from induction which learns an approximate probability generating function first and then uses it to evaluate a function at the points of interest, transductive inference estimates the values of a function at the points of interest in one step [15]. Hence this advantage of transductive learning helps adapt learning model to target data in a direct way.

In the context of VIS-NIR matching, we aim to perform VIS-NIR matching on the target people (Set A in Fig. 1) rather than the training set (Set B in Fig. 1), where target people do not have the labeled NIR mates in the training set, thus transduction can guide the matching learning adapted to them.

The difficulty of transductive VIS-NIR learning is that the distributions of gallery and probe sets are always not consistent in the original image space since images captured in VIS and NIR lighting conditions differ much from each other. As a result, the traditional transductive methods like KNN [9] and TSVM [12] could not be directly applied.

In this work, we overcome the above difficulty by developing 1) an illumination invariant feature representation (Sec. IV) and 2) an cross-domain transductive face matching model (Sec. V). The whole approach is illustrated in Fig. 2.

IV. Face Representation

A. Illumination Modeling

According to the Lambertian reflectance model, the intensity of a spectral face surface I can be represented as:

$$\begin{aligned} I(x,y) &= \rho(x,y)(\mathbf{n}(x,y)^T\mathbf{s}(x,y)) \\ &= |\mathbf{s}(x,y)|\rho_v(x,y)cos(\theta(x,y)), \end{aligned} \tag{1}$$

where R and L denote reflectance component and illumination component respectively, $\rho(x,y)$ is the albedo of the facial

surface material at point (x, y), $\mathbf{n}(x, y) = (n_x, n_y, n_z)^T$ is the surface normal (a unit vector) at surface point $z(x, y)$, $\mathbf{s}(x, y) = (s_x, s_y, s_z)^T$ is the point lighting direction and θ is the corresponding angle between $\mathbf{n}$ and $\mathbf{s}$.

We assume that the facial surface is Lambertian as in [6], [25], [28], and [32]. Then face imaging under VIS spot light and NIR near spot light can be modeled respectively by

$$I_v(x, y) = \rho_v(x, y)S_v(x, y), \tag{2}$$

and

$$I_n(x, y) = \rho_n(x, y)S_n(x, y), \tag{3}$$

where $S_v(x, y)$ and $S_n(x, y)$ are illumination patterns mainly caused by large scale objects and we denote them as large scale component [28], [29]. They are determined by the strength of light source and the angle between the light source direction and the surface normal. As shown in [6], [25], and [28], the large scale component is assumed to vary smoothly. Thus, within a small region around pixel (x_0, y_0), we assume $S_v(x, y) = c$ and $S_n(x, y) = c'$, for $(x, y) \in N_{(x_0,y_0)}$, where $N_{(x_0,y_0)}$ denotes the Neighborhood of pixel (x_0, y_0). That is, for each point $(x, y) \in N_{(x_0,y_0)}$, the Lambertians become

$$I_v(x, y) = c(x_0, y_0)\rho_v(x, y), \tag{4}$$

and

$$I_n(x, y) = c'(x_0, y_0)\rho_n(x, y), \tag{5}$$

where $c(x_0, y_0)$ and $c'(x_0, y_0)$ are constants related to $N_{(x_0,y_0)}$.

B. When and Why LBP and HOG Work?

Before introducing our proposed new descriptor, we would like to have a deep investigation of existing popular descriptors for VIS-NIR matching. Although LBP and HOG have been used in VIS-NIR face matching successfully [13], [18], it still lacks theoretical analysis on the illumination invariant property that tells why and when they can be used.

Note that the albedo components ρ_v for VIS and ρ_n for NIR are different. Nevertheless, since the surface of face is changing smoothly, the information within a small region will not vary too much. Hence, similar to [26], we draw the following assumption:

Assumption: within a small region of face image, ρ_v ***is approximately locally proportional to*** ρ_n***, that is,*** $\rho_v(x', y') = d(x, y)\rho_n(x', y')$ ***for some constant*** d ***that is determined by the neighborhood*** $N_{(x,y)}$**.**

In this paper, we additionally provide some statistical evidence to show the feasibility of this assumption as shown in Fig. 3. We have collected the block-wise cosine distances between pair-wise VIS-NIR images from all individuals and plotted the histogram in Fig. 3(d). The smaller the cosine distance is, the more linear proportional the relation between ρ_v and ρ_n is. As shown, most values are very small, indicating the approximately locally proportional relationship between I_v and I_n, which is also the relationship between ρ_v and ρ_n according to Eq. 4 and Eq. 5.

Now, we hold this assumption for the following analysis. The LBP encodes local pattern via thresholding the differences

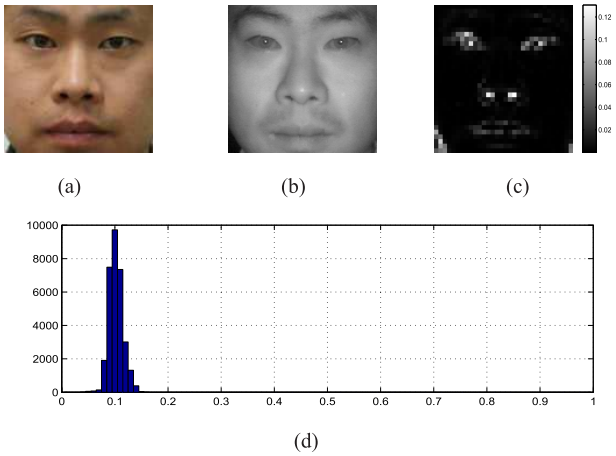

Fig. 3. Local cosine distance ($d_{cos}(x, y) = \frac{1}{\pi}acos(\frac{x^T y}{|x||y|})$) between corresponding VIS and NIR images. (a) VIS image; (b) NIR image; (c) block-wise local cosine distance between (a) and (b). We first divide the face into small blocks (with size of 4×4) and then calculate the cosine distance block-wisely between (a) and (b) and thus generate image (c). A small value in (c) indicates that the local vectors of (a) and (b) in the corresponding block are approximately collinear, i.e., being proportional. (d) Histogram of block-wise cosine distances for all pair-wiese VIS-NIR samples. Note that most values are small enough, where the $d_{cos}(x, y)$ is around 0.1 with variance less than 0.05.

between neighborhood pixels and the centered one, so that it is not sensitive to the local monotonous transformation. Since I_v and I_n are also approximately locally proportional, the local monotonous requirement is met. As a result, the LBP codings of a face image captured under VIS and NIR lighting conditions, i.e. $LBP(I_v)$ and $LBP(I_n)$, would probably be consistent under the assumption.

Similar observation holds for HOG, although the derivation is somehow different. HOG aims to encode an image by using gradient information, i.e., phase angle and local normalized gradient magnitude. According to [32], the phase angle has been proved to be invariant to homogeneous illumination change. It can be extended straightforwardly to our heterogeneous illumination situation as:

$$\frac{\partial_y I_v}{\partial_x I_v} = \frac{\partial_y \rho_v}{\partial_x \rho_v} \quad and \quad \frac{\partial_y I_n}{\partial_x I_n} = \frac{\partial_y \rho_n}{\partial_x \rho_n}, \tag{6}$$

where $\partial_x(\cdot)$ and $\partial_y(\cdot)$ are derivative operators along the x axis and y axis, respectively. Since ρ_v and ρ_n are approximately locally proportional, this leads to

$$\frac{\partial_y \rho_v}{\partial_x \rho_v} \approx \frac{\partial_y \rho_n}{\partial_x \rho_n}. \tag{7}$$

Thus we have

$$atan\left(\frac{\partial_y I_v}{\partial_x I_v}\right) \approx atan\left(\frac{\partial_y I_n}{\partial_x I_n}\right). \tag{8}$$

Besides, the local normalized gradient magnitude is computed by

$$\frac{\|\partial I_v(x, y)\|_2}{\sum\limits_{\substack{(x', y') \\ \in N_{(x,y)}}} \|\partial I_v(x', y')\|_2} = \frac{\|\partial \rho_v(x, y)\|_2}{\sum\limits_{\substack{(x', y') \\ \in N_{(x,y)}}} \|\partial \rho_v(x', y')\|_2}, \tag{9}$$

and

$$\frac{\|\partial I_n(x,y)\|_2}{\sum\limits_{\substack{(x',y')\\ \in N_{(x,y)}}} \|\partial I_n(x',y')\|_2} = \frac{\|\partial \rho_n(x,y)\|_2}{\sum\limits_{\substack{(x',y')\\ \in N_{(x,y)}}} \|\partial \rho_n(x',y')\|_2} \tag{10}$$

where $\partial I(x,y) = (\partial_x I(x,y), \partial_y I(x,y))$ and $\partial \rho(x,y) = (\partial_x \rho(x,y), \partial_y \rho(x,y))$.

Since ρ_v is proportional to ρ_n in $N_{(x,y)}$, it is easy to find that $\partial \rho_v$ is proportional to $\partial \rho_n$ in $N_{(x,y)}$ too. That is, $\partial \rho_v(x',y') = d(x,y)\partial \rho_n(x',y')$ for some constant d which is determined by $N_{(x,y)}$. We have

$$\frac{\|\partial I_v(x,y)\|_2}{\sum\limits_{\substack{(x',y')\\ \in N_{(x,y)}}} \|\partial I_v(x',y')\|_2} \approx \frac{\|\partial I_n(x,y)\|_2}{\sum\limits_{\substack{(x',y')\\ \in N_{(x,y)}}} \|\partial I_n(x',y')\|_2}. \tag{11}$$

Hence, according to Eq. (8) and Eq. (11), we find that the gradient histograms of VIS and NIR images from the same individual turn out to be the same, which makes LBP and HOG suitable for the VIS-NIR face matching.

C. Uniform Log-DoG Pattern Histogram

1) Log-DoG Pattern Histogram: Since shadow and noise are frequently presented in both VIS and NIR face images, directly applying LBP or HOG would probably lead to inaccurate description. Therefore, preprocessing is always taken before feature extraction. Liao et al. adopted Difference of Gaussian (DoG) to eliminate the difference between VIS and NIR images [18]; Yi et al. applied Laplacian of Gaussian (LoG) to encode the shared high frequent components in both VIS and NIR faces [30]. However, these methods are empirically designed and lack of theoretical analysis.

Based on the illumination model, we would like to derive a more robust representation by 1) taking the logarithm transformation to separate the identity related component ρ and large scale component S, 2) and then performing DoG filtering to alleviate the illumination effect and noise. For convenience, we denote this procedure as Log-DoG filtering. After that we encode the face pattern using the normalized local uniform pattern histogram.

As mentioned before, face image I can be represented as multiplication of the albedo ρ and the large scale component S in spatial domain. To better separate the albedo ρ and the large scale component S for further processing, we take the logarithm transform and make them combined additively. Note that, as indicated in [6], the large scale component S contains 1) background hues which always vary slowly and 2) important boundaries which are sharp edges, so S can be regarded as lying in the low or very high frequency domain. Consequently, as a kind of bandpass filtering methods, DoG suppresses both the lowest and highest spatial frequencies and thus is able to alleviate the effect caused by S, making modeling insensitive to noise in this scenario. As a result, the Log-DoG filtering procedure on I can be approximated by

$$\begin{aligned}&(G(x,y,\sigma_0) - G(x,y,\sigma_1)) * \log(I)\\ &\approx (G(x,y,\sigma_0) - G(x,y,\sigma_1)) * \log(\rho).\end{aligned} \tag{12}$$

After that, only partial information corresponding to mid-frequency part is kept. As indicated in [28], mid-frequency component in the logarithm domain is more likely related to facial structure rather than illumination effect. In the following, we will prove that the responses of LBP and HOG applied to such a component are also invariant to illumination change under the locally proportional assumption. As we know, the DoG image filtering is indeed a local linear operation, which is commutative to the derivation or division operation subject to the local patch. Therefore, without loss of generality, we only have to investigate the relationship of LBP/HOG codings between $\log(\rho_v)$ and $\log(\rho_n)$.

Remind that LBP is not sensitive to the local monotonous transformation and the logarithm operation is indeed exactly a kind of monotonous transformations. As a result, we get the following approximation based on the locally proportional assumption:

$$LBP(log(\rho_v)) \approx LBP(\rho_v) \approx LBP(\rho_n) \approx LBP(log(\rho_v)). \tag{13}$$

For the case of using HOG, the conclusion comes indirectly. Let us investigate the partial derivatives

$$\partial_x \log(\rho_v) = \frac{\partial_x \rho_v}{\rho_v}, \quad \partial_y \log(\rho_v) = \frac{\partial_y \rho_v}{\rho_v}, \tag{14}$$

and

$$\partial_x \log(\rho_n) = \frac{\partial_x \rho_n}{\rho_n}, \quad \partial_y \log(\rho_n) = \frac{\partial_y \rho_n}{\rho_n}. \tag{15}$$

Note that, Eq. (7) still holds in this situation, and gets similar conclusion as Eq. (8), since the factor $\frac{1}{\rho}$ can be eliminated during the division operation. By further substituting ∂I by $\partial \log(\rho)$, where $\partial \log(\rho) = (\partial_x \log(\rho), \partial_y \log(\rho))$, we can get the same right hand sides as those in Eq. (9) and Eq. (10) via Eq. (14) and Eq. (15). Thus, as indicated in Eq. (11) in subsection IV-B, the locally normalized gradient magnitude is also not sensitive to the illumination source. In summary, we can draw the conclusion that the gradient histograms of VIS and NIR images of the same instance after Log-DoG filtering are consistent under the locally proportional assumption.

Until now we have proved that it is feasible to extract domain invariant facial information by encoding the local patterns using LBP and HOG after Log-DoG filtering. In the rest of the paper, we denote these feature as LD-LBP and LD-HOG respectively, where LD is the abbreviation of "Log-DoG filtering". Some examples can be found in Fig. 4. Note that the LD-LBP images contain binary codings of the Log-DoG filtered images, and LD-HOG consists of two components, i.e., the local normalized gradient magnitude and gradient orientation. In addition, as shown in Fig. 5, we find that the Log-DoG filtering is able to alleviate the histograms difference between VIS and NIR domains.

2) Uniform Normalization: As discovered in our analysis, patterns/bins in feature coding are unevenly distributed as shown in Fig. 8(a)–(e). Some bins may rarely appear while some dominate in the same local patch, making the pattern histogram less informative and less compact, which may degrade the discriminant ability of the descriptor. Inspired by

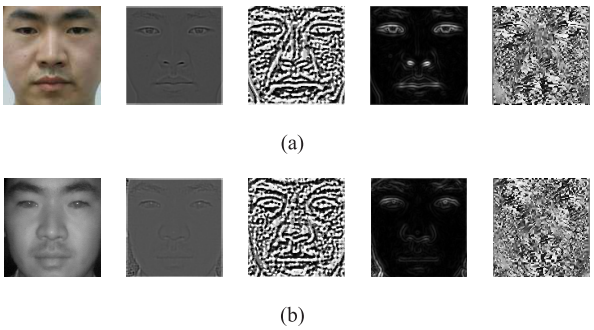

Fig. 4. Feature representation: from left to right are original image, results of proposed Log-DoG filtering (denoted as LD image), results of LD-LBP, gradient magnitude of LD image, orientation of LD image; the upper row are results of VIS image and the bottom row list results of NIR image.

related work on taking some normalization step for feature representation, e.g. merging the non 0−1 skipping bins taken by Ojala et al. [21] and uniformly distributed code histogram advised by Cao and Tang [5], we find that taking a feature normalization step during our transductive learning by using all available samples is of great help for the latter feature quantification. The normalized strategy we used here is:

$$f_{norm}(i) = \frac{f(i) - mean(\mathbf{f^i})}{std(\mathbf{f^i})}, \tag{16}$$

where f_{norm} is the normalized feature of f, and $\mathbf{f^i}$ denotes the array constituted by the i^{th} bin of features from all faces.

Note that, as illustrated later in Fig. 8(c) and (f), the main characteristics of the uniform normalization can be summarized as follow.

1) There are both labeled and unlabeled entries in the transductive framework, providing more accurate description on the feature distributions when taking the normalization step;
2) All bins are evenly distributed after normalization, providing unprejudiced opportunity for each bin in the next feature extraction step;
3) It takes samples from both VIS and NIR domains into account, good for filing the domain gap.

V. Transductive VIS-NIR Face Matching

Though the aforementioned LD-LBP and LD-HOG feature representations are designed to minimize the gap between VIS and NIR domains, it is not enough to tackle the cross-domain matching problem well, because the Lambertian assumption or the locally proportional assumption may not strictly hold. Hence, we wish to further quantify our derived features using a proposed transductive subspace model called Transductive Heterogeneous Face Matching (THFM) so as to further pursue extraction of invariant features for VIS-NIR matching. We elaborate our proposed model in four steps: domain invariant feature extraction, target related discriminant model learning, cross domain penalization and locality preserving. Through our modeling, we formulate a transductive framework to match probe NIR images to the gallery VIS images (Set A in Fig. 1) assisted by the training (Set B in Fig. 1), which only holds labeled VIS-NIR pairs during learning.

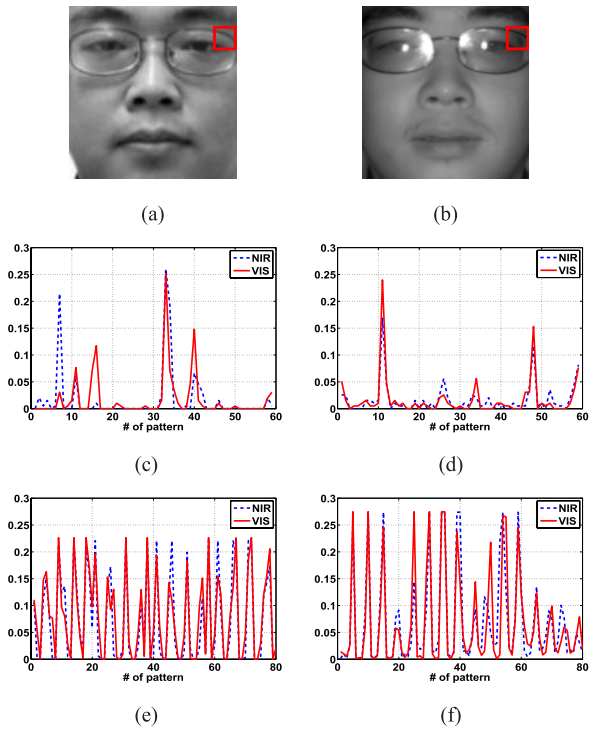

Fig. 5. Local encoding with and without Log-DoG filtering. (a) VIS face; (b) NIR face; (c) local patch LBP coding; (d) local patch LD-LBP coding; (e) local patch HOG coding; (f) local patch LD-HOG coding.

A. Domain Invariant Feature Extraction

In order to seek the domain invariant feature or shared features in VIS and NIR domains, we further extract domain invariant components based on the proposed representation in a latent subspace. Since there are available labeled training VIS-NIR images (Set B in Fig. 1) for us to investigate the relationship between two domains, we hope that the expected feature mapping $\phi(x) : \mathcal{X} \rightarrow \mathcal{H}$ could draw points of the same class together no matter which domain they belong to. We achieve it by minimizing the intra-class variation of samples of training people in the feature space.

Suppose we have labeled VIS-NIR images $X^S = [x^S_{q,k}]$, $\{x^S_{q,k} | q \in C_S\} = \{x^{S\_VIS}_{q,i} | q \in C_S\} \cup \{x^{S\_NIR}_{q,j} | q \in C_S\}$, where $x^S_{q,k}$ is the k^{th} sample of the q^{th} training people. The intra-class variation in the latent feature space $\mathcal{H}$ can be quantified by the trace of average intra-class scatter matrix S_{intra}:

$$S_{intra} = \frac{1}{N_S} \sum_{k,q \in C_S} \left(\phi(x^S_{q,k}) - \bar{m}^S_q\right) \left(\phi(x^S_{q,k}) - \bar{m}^S_q\right)^T, \tag{17}$$

where $\bar{m}^S_q$ is the mean vector of the q^{th} training people in the latent space $\mathcal{H}$, and N_S denotes the total amount of samples from the training.

Let $X = \left[X^S, X^{Gallery}, X^{Probe}\right]$, kernel matrix $K = \Phi(X)^T \Phi(X)$, where $X^{Gallery}$ and X^{Probe} denote the gallery VIS set and probe NIR set defined in Section III-A respectively, and $\Phi(X) = [\phi(x)]_{x \in X}$. Our objective is to address the

minimization problem below:

$$\min_{\phi} tr(S_{intra}) = \frac{1}{N_S}\sum_{k,q\in C_S}\left(\phi(x_{q,k}^S)-\bar{m}_q^S\right)\left(\phi(x_{q,k}^S)-\bar{m}_q^S\right)^T = tr\left(\tilde{S}_{intra}K\right), \quad (18)$$

where $\tilde{S}_{intra} = \begin{bmatrix} \frac{I_{N_S}-\sum A_k A_k^T}{N_S} & 0\ 0 \\ 0 & 0\ 0 \\ 0 & 0\ 0 \end{bmatrix}$,

$A_k = [a_{kj}]_{1\times N_S}$, $a_{kj} = \begin{cases} \frac{1}{\sqrt{N_q}}, & if\ label(x_j^S) = k \\ 0, & otherwise \end{cases}$.

B. Target Related Discriminant Feature Mining

Matching probe images to the gallery images (Set A in Fig. 1) can be seen as applying a classifier trained on the gallery set to those probe images. However, it works only if the samples in the gallery and probe set are coming from the same domain. Fortunately, as we consider domain invariant feature extraction and classification on target people simultaneously under the transductive framework, we are able to cope with the heterogeneous matching problem in a homogeneous way. That is, the discriminant model trained by gallery set would be as good as possible to fit the probe images in the latent space. Thus, we hope that the discriminant information of gallery individuals could be preserved in the desired space. There are various works addressing the discriminant feature mining problem, such as FDA [1], MMC [16], SVM [8] etc. In this work, we adopt the between-class variation to measure the separability of individuals in gallery set, which is suitable for subspace analysis and can be well described by the inter-class scatter matrix and proved to be effective in FDA. Using the notation in Section III, our objective is to maximize the trace of the average between-class variation in gallery set as follow:

$$\max_{\phi} tr(S_{inter}) = tr\left(\frac{1}{N_G}\sum_{i\in C_T} N_i^G(\bar{m}_i^G-\bar{m}^G)(\bar{m}_i^G-\bar{m}^G)^T\right) = tr\left(\tilde{S}_{inter}K\right), \quad (19)$$

where N_i^G is the number of samples in class i, $N_G = \sum_i N_i^G$, $\bar{m}_i^G$ denotes the mean of class i and $\bar{m}^G$ is the mean of all galleries. $\tilde{S}_{inter} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{\sum N_i^G B_i B_i^T}{N_G} & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $B_i = [b_{ij}]_{1\times N^G}$,

$b_{ij} = \begin{cases} \frac{1}{N_i^G} - \frac{1}{N_G}, & if\ label\left(x_j^{Gallery}\right) = i \\ -\frac{1}{N_G}, & otherwise \end{cases}$.

C. Cross Domain Penalization

In Section V-A, we have considered the relation between VIS and NIR domains for training people (Set B in Fig. 1). However, since our goal is to match target people (Set A in Fig. 1), we also have to guarantee that the VIS and NIR images of the same target people have the same or similar representation in the feature space. However, albeit conducting transductive learning, the label information of probe images is unknown, which makes it impossible to align the VIS and NIR face images via label information.

We attempt to minimize the Maximum Mean Discrepancy (MMD) [3] for penalizing the cross domain difference between gallery and probe sets for target people. Consequently, given samples $X^G = \{x_{p,i}^{Gallery} | p \in C_T\}$ and $X^P = \{x_{p,i}^{Probe} | p \in C_T\}$ drawn from two different domains and the kernel-induced feature map ϕ, the estimate of MMD between X^G and X^P in the feature space is as follows:

$$MMD(\Phi(X^G), \Phi(X^P)) = \left\| \frac{1}{N_G}\sum_{p,i}\phi\left(x_{p,i}^{Gallery}\right) - \frac{1}{N_P}\sum_{p,j}\phi\left(x_{p,j}^{Probe}\right)\right\|_{\mathcal{H}}^2. \quad (20)$$

The $MMD\left(\Phi\left(X^G\right), \Phi\left(X^P\right)\right)$ can be written as:

$$MMD\left(\Phi\left(X^G\right), \Phi\left(X^P\right)\right) = tr(MK) \quad (21)$$

where $M = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{\mathbf{1}_{N_G}\mathbf{1}_{N_G}^T}{{N_G}^2} & -\frac{\mathbf{1}_{N_G}\mathbf{1}_{N_P}^T}{N_G N_P} \\ 0 & -\frac{\mathbf{1}_{N_P}\mathbf{1}_{N_G}^T}{N_G N_P} & \frac{\mathbf{1}_{N_P}\mathbf{1}_{N_P}^T}{N_P N_P} \end{bmatrix}$, $\mathbf{1}_{N_G} \in \mathbb{R}^{N_G}$ and $\mathbf{1}_{N_P} \in \mathbb{R}^{N_P}$ denotes the column vector with all 1's, N_G and N_P are the total amount of samples in X^G and X^P, respectively.

D. Locality Preserving

In order to preserve the structure of data as much as possible during domain adaptation, we share the intrinsic locality property of the manifold regularizer [2] via constructing a graph with the affinity $v_{ij} = 1$ if x_i is the k nearest neighbors of x_j, or vice versa; otherwise $v_{ij} = 0$. Let $V = [v_{ij}]$, and the graph Laplacian matrix becomes $\mathcal{L} = D - V$, where D is the diagonal matrix with entries $d_{ii} = \sum_{j=1}^n v_{ij}$. Note that, the embedding coordinate of x_i in $\mathcal{H}$ is $\phi(x_i)$. To maintain the neighborhood structure, we absorb the manifold learning idea [2] and formulate the following objective for minimization

$$\sum_{(i,j)\in\mathcal{N}} v_{ij}\left|\left|\phi(x_i) - \phi(x_j)\right|\right|^2 = tr(\mathcal{L}K) \quad (22)$$

E. Transductive Heterogeneous Face Matching (THFM) Framework

Putting all together, we attempt to learn a latent feature space $\mathcal{H}$ in which the four aforementioned objectives from Sec. V-A to Sec. V-D can be optimized simultaneously. However, directly addressing kernel matrix K is expensive and cannot handle unseen patterns as well. In order to tackle this problem, we adopt the unified kernel learning method used in [22] to learn a low dimensional feature space.

Note that, according to the empirical kernel map [23], the kernel matrix K can be decomposed as $K = (KK^{-1/2})(K^{-1/2}K)$. Consider the use of a matrix $\tilde{W} \in \mathbb{R}^{N\times m}$ that transforms the empirical kernel map to an m-dimensional space (where $m \ll N$). We then obtain the following kernel matrix

$$\tilde{K} = (KK^{-1/2}\tilde{W})(\tilde{W}K^{-1/2}K) = KWW^TK, \quad (23)$$

where $W = K^{-1/2}\tilde{W}$. Intuitively, the kernel evaluation between any two patterns x_i and x_j is $\tilde{k}(x_i, x_j) = k_{x_i}^T W W^T k_{x_i}$, where $k_x = \left[k(x_1, x), ..., k(x_N, x)\right]^T$. That is, this kernel $\tilde{k}$ is able to facilitate the out-of-sample kernel evaluations.

By substituting the kernel matrix $\tilde{K}$ into Eq. (18, 19, 21, 22), we develop our objective function in the form of Rayleigh quotient as follows:

$$\max_{W \in \mathbb{R}^{N \times d}} \frac{tr(W^T K \tilde{S}_{inter} K W)}{tr(W^T (K(\tilde{S}_{intra} + \beta M + \gamma L)K + \alpha I)W)}, \quad (24)$$

where $\beta \geq 0$ and $\gamma \geq 0$ are the hyper-parameters for the MMD term and manifold regularized term respectively, and α is the weight for Tikhonov regularization which is used to avoid degeneration in the generalized eigen-decomposition problem. Finally, the solution to Eq. (24) is equal to the leading eigenvectors computed by the following general eigenvalue problem

$$K \tilde{S}_{inter} K w = \lambda (K(\tilde{S}_{intra} + \beta M + \gamma L)K + \alpha I)w, \quad (25)$$

where w is the eigenvector and λ denotes the corresponding eigenvalue. Note that, only eigenvectors with non-zero eigenvalue are used in our final result.

VI. Experiments

A. Dataset

We conduct evaluations on the following two public datasets:

Surrey's VIS-NIR Database [27]: It consists of 1056 visible light face images and 1056 Near-IR faces from 22 individuals, each with 48 pairs of VIS and NIR images captured at the same time. Note that, there are 4 different lighting directions in the visible light images: top, left, bottom and right. The Near-Infrared images are taken under frontal Near-IR flash illumination. All images are manually aligned according to the eye coordinates, cropped and resized to 128×128 pixels.

CASIA-HFB: It is a popular heterogeneous face database widely used in [13], [14]. We only select the subset containing VIS-NIR images. There are 2095 VIS images and 3002 NIR images from 202 individuals. Note that, all faces are frontal and there are some variations including expression, wearing glasses, scaling, etc. They are manually aligned according to the eye coordinates, cropped and resized to 128×128 pixels as shown in Fig. 1.

B. Experimental Setting

In our experimental validations, we aim to evaluate (i) whether the feature we used can alleviate the domain difference and retain discriminative information for VIS-NIR face matching; (ii) whether the proposed learning method is effective in mining useful information for such a cross domain face matching/recognition.

We evaluate various descriptors in our experiments, including original LBP [21], HOG [10] and their combination (denoted as LBP-HOG), our proposed descriptors, including LD-LBP, LD-HOG, LD-LBP-HOG and the uniform normalization based ones (denoted as NLBP, NHOG, NLBP-HOG, NLD-LBP, NLD-HOG, NLD-LBP-HOG) as shown in Sec. IV-C.2. More specifically, for each face image with 128×128 pixels, we firstly divide it into 8×8 blocks with size of 16×16 and then encode the local pattern for each block using the above descriptors respectively. Note that, 59 bin histogram is used to form LBP coding and the parameters of HOG are listed in Table I. Hence, the feature dimensions for each image are 3776 for LBP based features and 5120 for HOG based features.

TABLE I
Default Parameter Setting for Our Methods

Method	Parameter	Value
DoG	σ_0	1
	σ_1	2
HOG	block size	16×16
	cell size	4×4
	issigned	0
	bins in cell	5
THFM	kernel type	linear
	α	0.01
	β	0.001
	γ	0.1
	k nearest neighbors in L	3

In addition, we take Tan and Triggs filter (denoted as TT) based features as baselines for comparison, denoted as TT-LBP, TT-HOG and TT-LBP-HOG, where for example TT-LBP means TT filtering followed by LBP coding.

We compare the proposed THFM to classical VIS-NIR matching approaches, such as PCA, Fisher Discriminant Analysis (FDA) [1], Canonical Correlation Analysis (CCA) based methods [31] and Coupled Spectral Regression (CSR) [14]. Meanwhile, the results of state-of-the-art algorithms, such as Local Structures of Normalized Appearance (LSNA) [18], Random Subspace [13], Light Source Invariant Features (LSIFs) [20] are also reported for comparison. However, neither of them can perform transductively. To be fair, we compare THFM with related subspace methods such as transductive PCA (tPCA), TCA [22], SDA[1] [4] for performing VIS-NIR matching in a transductive way, although THFM is a first transductive model for such a heterogeneous face matching.

Unless otherwise specified, the parameters of our proposed THFM are set as those in Table I. Note that, we have tried linear, RBF and Polynomial kernels in our experiments and found that the results of using linear kernel always turn out to be the best, which was similarly reported in [22] for domain adaptation. The reason might be that the dimensionality of the feature is higher than the number of sample, so it's possible that the samples may be almost linearly separable. Note that using linear kernel is equivalent to letting ϕ be the empirical kernel in terms of linear projection. Further discussion about the parameter selection could be found in Section VI-E. For the compared methods, all parameters are tuned carefully to obtain the best performance. For PCA related approaches, the ratio of principal components is set to 95%. For CSR, λ and η

[1] In some case, semi-supervised learning can be also regarded as transductive learning [4].

TABLE II
AVERAGE CLASSIFICATION ACCURACY (%) OF VARIOUS METHODS ON SURREY'S VIS-NIR DATABASE
(THE NUMBER INSIDE PARENTHESES IS THE STANDARD DEVIATION)

	NoLearing	**PCA**	**FDA**	**PCA+CCA**	**PCA+CCA**	**CSR**	**tPCA**	**TCA**	**SDA**	**THFM**
Raw	8.91(3.31)	14.28(3.90)	8.65(1.89)	6.84(2.81)	6.39(2.63)	8.74(2.07)	64.80(2.09)	68.31(2.29)	76.39(1.71)	72.70(5.87)
LBP	61.67(2.67)	50.75(5.85)	9.17(1.59)	7.58(2.45)	8.15(2.01)	6.40(2.00)	79.59(2.34)	71.84(2.62)	99.59(0.27)	98.58(1.38)
TT-LBP	84.06(1.70)	54.65(5.62)	7.77(2.34)	7.08(1.20)	7.91(2.42)	5.00(0)	87.08(1.38)	74.71(1.29)	95.54(0.68)	98.66(0.66)
LD-LBP	84.15(1.55)	61.21(4.34)	7.72(2.02)	7.59(1.38)	7.99(2.31)	5.00(0)	90.21(1.02)	78.63(1.12)	97.93(0.47)	95.84(1.22)
NLD-LBP	94.56(0.54)	24.12(1.36)	8.96(0.97)	7.49(0.83)	7.85(1.69)	8.47(1.02)	80.44(1.25)	84.40(0.99)	94.94(0.68)	99.14(0.12)
HOG	88.17(1.33)	52.10(8.40)	9.38(1.58)	10.78(3.13)	8.43(2.26)	8.75(1.70)	87.78(1.39)	87.97(1.36)	98.47(1.32)	99.93(0.10)
TT-HOG	94.20(1.54)	51.54(5.87)	8.50(1.86)	8.82(1.90)	8.64(1.87)	7.48(2.09)	91.86(0.76)	91.76(0.74)	98.84(0.18)	100(0)
LD-HOG	90.05(1.86)	61.33(4.25)	8.13(1.80)	6.08(1.54)	8.30(1.83)	6.69(2.28)	90.31(0.84)	90.33(0.92)	98.40(0.29)	99.78(0.08)
NLD-HOG	93.02(1.86)	61.34(3.59)	9.24(0.54)	7.67(1.23)	9.61(0.90)	9.20(0.71)	96.80(0.50)	97.15(0.42)	99.88(0.11)	100(0)
LBP-HOG	82.46(1.85)	51.08(7.71)	9.23(1.62)	10.74(2.73)	9.07(1.57)	8.88(1.85)	84.23(1.71)	86.42(1.41)	99.51(0.44)	99.97(0.05)
TT-LBP-HOG	91.60(0.67)	51.79(6.12)	8.55(1.87)	8.85(2.33)	8.57(1.92)	7.70(2.00)	91.87(0.77)	94.22(1.49)	98.94(0.21)	100(0)
LD-LBP-HOG	90.66(0.83)	60.60(4.39)	8.26(1.79)	6.54(1.59)	7.99(2.07)	6.65(2.31)	90.60(0.77)	90.43(1.86)	98.99(0.15)	99.83(0.07)
NLD-LBP-HOG	97.60(0.37)	46.22(2.93)	9.51(0.38)	7.54(1.45)	9.60(0.56)	9.27(1.11)	97.45(0.39)	94.77(1.80)	99.88(0.11)	**100(0)**

*Note: the results of THFM and those using our proposed features are marked with gray background. See text for detailed setting.

are set as 1 and 0.01 respectively. Besides, $\alpha = 0.1$ and $dim = c-1$ are used in TCA [22] where c denoted the number of classes in gallery set, and α and β in SDA [4] are set to 0.1 and 0.001. For all methods, cosine distance and nearest neighbor classifier are used to compute final recognition rate.

C. Surrey's VIS-NIR Database

We would like to investigate the performances of different methods when there are only a few training samples available. Here, we selected the Surrey database for the evaluation. Only 2 individuals were randomly selected for training, i.e., only VIS and NIR images of these 2 individuals had labels, and the remaining 20 individuals were considered as probe ones. To make the evaluation more challenging, we only selected 4 images (one in each lighting condition) for each individual to construct the gallery, and finally there were 4 VIS images and 48 probe NIR images per individual in the testing procedure. We tried Raw image, LBP, HOG and the proposed local encoding features, i.e. LD-LBP, NLD-LBP, LD-HOG and NLD-HOG. Considering that TT filtering is able to removal sub-surface scattering, we conducted TT-LBP and TT-HOG as baselines. Finally, we repeated the experiment 10 times and report the mean accuracies and standard deviations.

As shown in the first column of Table II, without applying any learning methods, our proposed NLD-LBP and NLD-HOG almost outperform all other descriptors except TT-HOG. Note that the TT based features are also powerful in alleviating heterogeneous difference caused by VIS and NIR lighting, achieving comparable results to some of our proposed methods in Surreys dataset. However, we note that although TT and Log-DoG (LD) are all for pre-processing, the rationale of our proposed descriptor for VIS-NIR matching is supported by theoretical analysis (see Sec. IV-C.1), while TT still lacks of strong theoretical support for such a cross-domain matching.

When learning is further incorporated, we find that the transductive methods, especially SDA and our proposed THFM, perform substantially better than the non-transductive algorithms and achieve extremely high accuracies in this challenging setting where only limited labeled VIS-NIR pairwise training samples provided. On the other hand, the non-transductive algorithms such as FDA, CCA based methods and CSR, suffer from severe degeneration, with average accuracies less than 10%. The reason might be that those discriminant methods focus on discriminative components only for the limited training people, which may not be suitable for probe ones, while the transductive ones are able to fill the gap between training and target people much better. Note that SDA is not optimal for VIS-NIR face matching problem, although it performs as well as THFM in some cases. To see the superiority of our proposed THFM, a more challenging database will be used in the next section. Overall, our proposed THFM almost achieves the best performances.

D. CASIA-HFB Dataset

In this section, we conducted a series of experiments to perform the evaluation on a more large-scale and challenging **CASIA-HFB** database. By following Klare and Jain's setting in [13] where extensive evaluations have been reported, 100 individuals were selected as training people and 100 individual were selected as the target ones (including the gallery images and probe images). The general results can be found in Table III. Note that standard deviation is not available in Table III due to the testing protocol in [13]. It is interesting to find that, without incorporating any learning method, the TT based features and our proposed descriptors achieve comparable results, which are much better than those of original LBP and HOG. But the difference becomes clear when learning methods are involved. In most of the cases, our proposed features, especially NLD-LBP-HOG, always outperform other descriptors no matter using our proposed THFM or using other learning methods.

1) Comparison With Classical Methods: As shown in Fig. 6(a) and (c), we can see that, the Log-DoG filtering procedure eliminates domain difference a lot since the performances of PCA are substantially boosted after the processing. What's more, both FDA and THFM that take the data structure information into account can maintain or improve their performances by using Log-DoG based features. In comparison, the

TABLE III
CLASSIFICATION ACCURACY (%) OF VARIOUS METHODS ON CISIA-HFB DATABASE

	NoLearing	PCA	FDA	PCA+CCA	PCA+CCA	CSR	tPCA	TCA	SDA	THFM
Raw	3.16	3.16	53.53	26.70	39.81	38.92	3.16	0.21	38.30	20.04
LBP	4.32	5.06	16.61	27.61	18.11	6.44	5.22	12.39	40.83	32.94
TT-LBP	42.42	40.63	11.46	15.79	11.26	0.69	43.58	44.89	66.23	46.26
LD-LBP	45.37	46.44	18.44	20.11	16.89	0.78	49.83	52.22	56.78	44.83
NLD-LBP	41.94	31.11	19.50	17.72	16.67	23.17	41.22	46.67	29.89	71.28
HOG	6.59	5.39	53.50	49.67	48.83	10.17	7.33	21.61	51.94	62.00
TT-HOG	77.21	62.94	51.00	38.09	30.47	11.26	77.90	61.15	73.99	73.71
LD-HOG	77.69	64.50	55.44	44.94	33.22	27.22	80.11	68.50	61.72	70.50
NLD-HOG	81.67	66.83	66.67	44.44	37.94	47.89	81.39	70.72	83.06	98.72
LBP-HOG	4.53	5.28	61.22	54.67	54.17	13.67	6.00	22.67	49.78	58.89
TT-LBP-HOG	76.87	62.46	62.73	40.77	34.80	17.09	77.69	60.74	76.66	73.51
LD-LBP-HOG	77.63	64.61	62.00	47.06	36.61	21.61	80.33	68.56	65.00	70.39
NLD-LBP-HOG	77.21	53.33	76.33	48.56	46.89	54.89	76.28	70.06	85.00	**99.28**

*Note: the results of THFM and those using our proposed features are marked with gray background. See text for detailed setting.

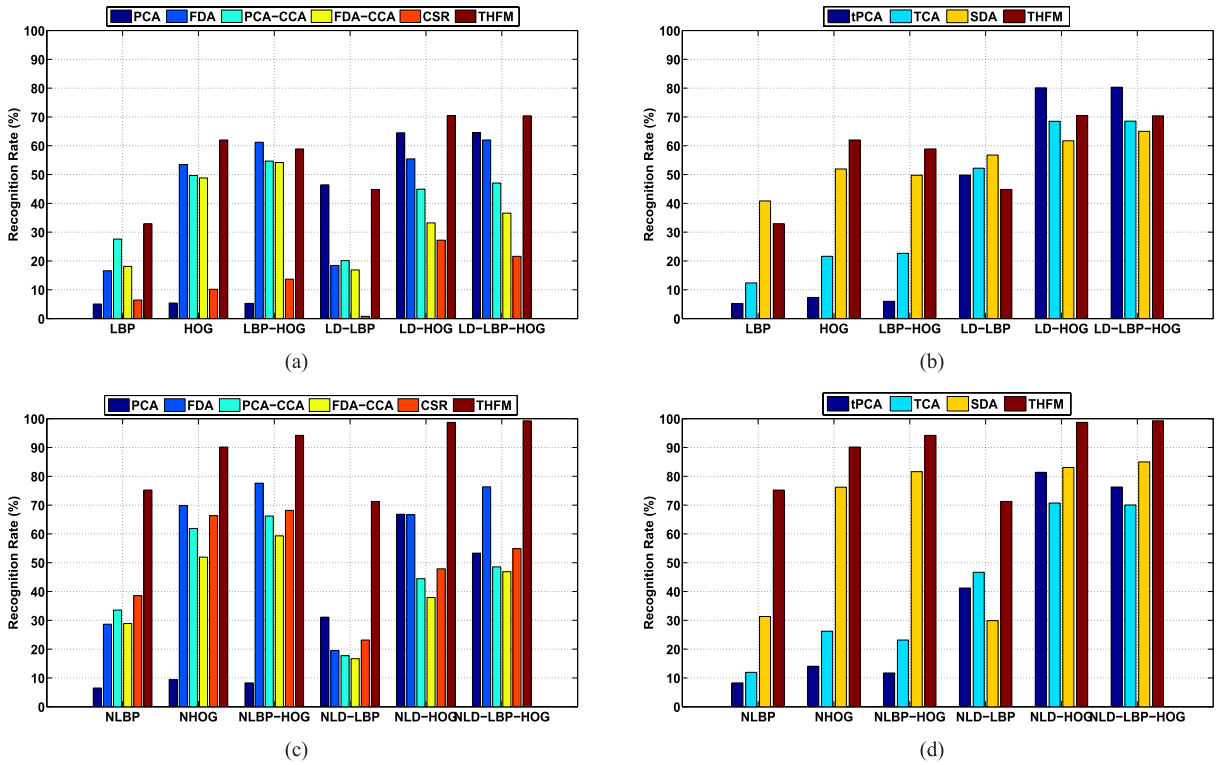

Fig. 6. Comparison of different methods using different features on the CASIA-HFB database. (a) and (c) compare THFM to non-transductive methods: PCA, FDA, PCA+CCA, FDA+CCA, CSR; (b) and (d) compare THFM to transductive methods: tPCA, TCA, SDA, THFM. (a) and (b) are based on the non-normalized features, while (c) and (d) are based on the normalized features.

CCA based methods, which only consider the discriminative information for training but ignore the difference of data distributions between training and target, easily tend to be overfitting and thus do not perform well.

When formulating the VIS-NIR face matching problem in a transductive framework and extracting valuable information according to the objective criterion, our proposed THFM substantially outperforms those methods no matter which type of feature is used. An impressive 99.28% rank-1 recognition rate is gained by combining NLD-LBP and NLD-HOG features.

2) Comparison With Transductive Methods: As shown in Fig. 6(b) and (d), The performances of THFM with normalized features are of the best among those of all four approaches, i.e. tPCA, TCA, SDA and THFM. On one hand, our proposed THFM and SDA outperform TCA and tPCA in most of the cases, suggesting that supervised information is indeed useful for cross domain matching; on the other hand, THFM is better than SDA by 10% on average. The reason could be that SDA is only optimal when the probe set and gallery set are coming from the same domain, which is different from the VIS-NIR heterogeneous setting. In comparison, THFM is able to handle the cross domain matching problem more effectively, by minimizing the domain difference and retaining the discriminative information for classification on target people at the same time.

TABLE IV
CLASSIFICATION ACCURACIES OF VARIOUS STATE-OF-THE-ART METHODS

	FAR=0.1%	FAR=1%	Rank-1
THFM	**98.42%**	**99.66%**	**99.28%**
Goswami's [11]	50.64%	75.55%	83.28%
Liao's* [18]	68.00%	87.50%	-
NNSR [13]	79.05%	91.37%	90.38%
FV [13]	85.62%	93.80%	97.60%
NNSR+FV [13]	93.45%	97.06%	97.63%
Yi's* [30]	92.00%	98.00%	91.67%
Liu's* [20]	90.00%	98.00%	98.51%

* Results of these three approaches are reported in [20], where 150 individuals are used for learning and the other 50 individuals consist the testing set.

It is worth mentioning that, though tPCA and TCA are unsupervised methods, they also obtain surprising results by using Log-DoG based features, especially those encoded by HOG. In other words, it indicates that the Log-DoG based features, especially LD-HOG, are good for the VIS-NIR face matching problem on such a more challenging CASIA-HFB database.

3) Comparison With State-of-the-Art Methods: We also compare our method with the state-of-the-art methods in this field, including Goswami's LDA using TT [24] based feature [11] combined with LBP+HOG, Liao's LSNA [18], Klare's feature ensemble approach using both nearest neighbor and sparse representation matching [13], the baseline commercial face recognition system FaceVACS used in [13], Yi's binary LoG local matching strategy [30] and the multi-DoG filtering followed by Adaboost feature selection proposed by Liu [20]. Note that, only 50 individuals are set as target in Liao's and Liu's works.

In Table IV, three performance measures are reported for comparison, including Verification Rate (VR) at False Accept Rate (FAR) $= 0.1\%$, VR at FAR $= 1\%$ and Rank-1 recognition rate. It should be noted that most of the results are excerpted from [13] and [20], and the Rank-1 recognition rates of Liao's LSNA is not available. For our approach, the normalized LD-LBP-HOG is used as the feature representation.

As shown in Table IV, our proposed method outperforms all other state-of-the-art methods in all the three different indexes, achieving VR $= 98.42\%$ at FAR $= 0.1\%$, VR $= 99.66\%$ at FAR $= 1\%$ and 99.28% Rank-1 accuracy. Compared to those local feature based methods, the advantages of our proposed method are in two folds: (i) different from [18] and [20], we do not implement adaboost to select discriminative feature but perform feature normalization instead, which is efficient enough and able to maintain data distribution structure due to the unsupervised feature representation; (ii) Instead of learning the inductive model for VIS-NIR face matching, we consider the matching problem in a transductive framework. More specifically, our proposed method learns the VIS-NIR face matching assisted on a small set of VIS-NIR samples from training and adapt it for target people through transductive learning.

E. Further Investigation

In this section, we would like to compare the effectiveness of different features and investigate what properties are more favourable for our VIS-NIR face matching model. The **CASIA-HFB** database is used as the benchmark.

Normalized vs. Non-Normalized Features. Note that our normalization processing is independent of the type of feature and can be applied to all features. We tested FDA, SDA, tPCA, and the proposed THFM by using LBP, HOG, LBP-HOG, LD-LBP, LD-HOG, LD-LBP-HOG and their corresponding normalized forms, i.e., NLBP, NHOG, NLBP-HOG, NLD-LBP, NLD-HOG, NLD-LBP-HOG. The results of different methods conducted based on the normalized feature are denoted as FDA(N), SDA(N), tPCA(N) and THFM(N). To better investigate the effectiveness of this normalization step for cross domain face matching, we report the ROC curves of each algorithm using features with and without normalization. All results are illustrated in Fig. 7, where the results with normalization processing are denoted by solid line while those without normalization processing are denoted by dotted line.

As shown, there are always notable improvement when using normalized features, i.e. NLD-LBP, NLD-HOG, NLD-LBP-HOG. Note that tPCA does not get improved clearly. The reason might be that it removes the second-order statistical relationship among different dimensions, which can be considered as a kind of implicit feature normalization by transforming the original feature distribution to a sphere form.

In order to further investigate the effect of normalization, we compare the variance of feature for each dimension with and without normalization. 20 bins of various features were randomly chosen and their distributions in VIS and NIR domains are plotted separately in Fig. 8. We find that the feature distributions between VIS and NIR domains are similar due to domain invariant property, supporting the analysis in Sec IV. Also, as shown in Fig. 8(c) and (f), using NLD-LBP and NLD-HOG would make the distributions of features more invariant across domains and thus meet the requirement of domain adaptation.

Sensitivity to Parameter Selection. We evaluate THFM for different parameter settings using the **CASIA-HFB** database. We use linear kernel and the dimensionality of final latent space was fixed to be $c - 1$, where c is the amount of classes in gallery set. The other default parameters are listed in Table I. For evaluation here, we always fix other parameters and vary the one we are concerned about. Note that, all results are based on the NLD-LBP-HOG feature which is proved to be the most effective for our proposed THFM in the reported experiments.

As shown in Fig. 9(a), a small α is more suitable since it is used to avoid degeneration. Thus, we would like to set it as a small constant, e.g., $\alpha = 0.01$ in our setting. Regarding β, the performance drops dramatically when β is larger than 0.1. However, it performs well when β lies in the range $[0.001, 0.1]$. Note that, the parameters γ and k are used for the Laplacian regularization. Fig. 9(c) illustrates the variations of our method for different γ, and one can find the peak of the performance around $\gamma = 1$. Finally, we find that THFM

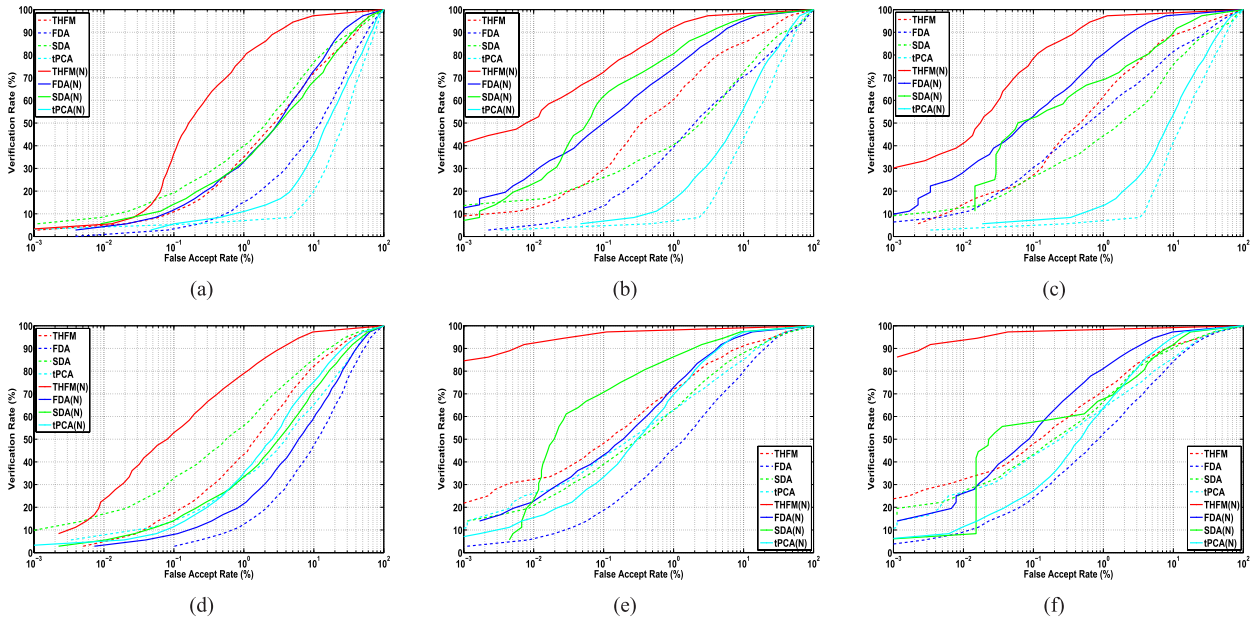

Fig. 7. ROC curves of various algorithms using with (marked with solid line) and without (marked with dashed line) normalized features: (a) LBP; (b) HOG; (c) LBP+HOG; (d) LD-LBP; (e) LD-HOG; (f) LD-LBP-HOG

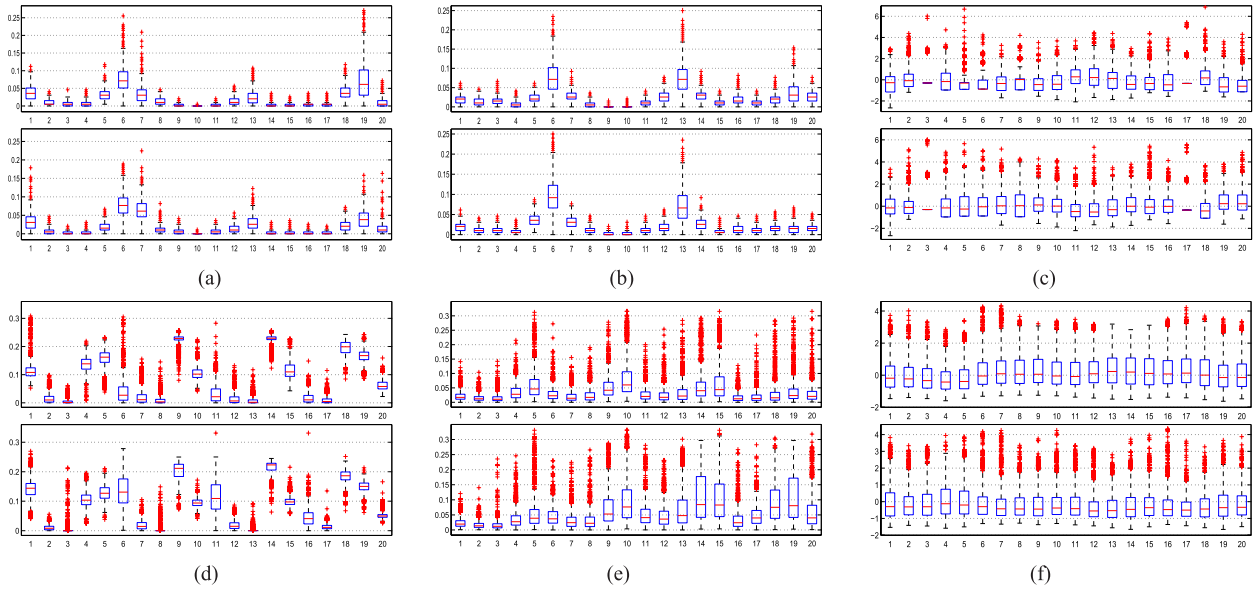

Fig. 8. Distributions of different features in VIS and NIR domains: (a) LBP; (b) LD-LBP; (c) NLD-LBP; (d) HOG; (e) LD-HOG; (f) NLD-HOG. Box plot is used here to describe the distributions, and results of VIS domain are shown in upper rows, while those of NIR domain are shown in lower rows.

is less sensitive to k and the performances of the method for different k are ranging from 98.5% to 99.5%.

Kernels. We also evaluate the performance of our method using different kernels. Besides linear kernel, RBF kernel and polynomial kernel (d = 3) are also investigated in our experiments. Note that, we have turned the parameters for each kernel to ensure that the best performances are gained and the results are shown in Table V. As shown, linear kernel is good enough for our cross domain VIS-NIR face matching and performs better than the other two kernels. This agrees with the well-known observation that the linear kernel is often adequate for high dimensional data [22].

Out of Sample. Although our proposed THFM is designed in a transductive framework, it is also able to address the out-of-sample problem in a single domain. In this part, following the setting in VI-D, we conducted an out-of-sample evaluation experiment by dividing the probe NIR images into two parts, half for learning and half for evaluating. More specifically, 920 probing NIR images from 100 testing individuals together with training and gallery images were used to learn the kernel evaluation function k_x, and then we tested on the rest 880 NIR images which were not involved in the learning procedure. The final Rank-1 accuracy of using NLD-LBP-HOG feature representation is 98.86%, almost the

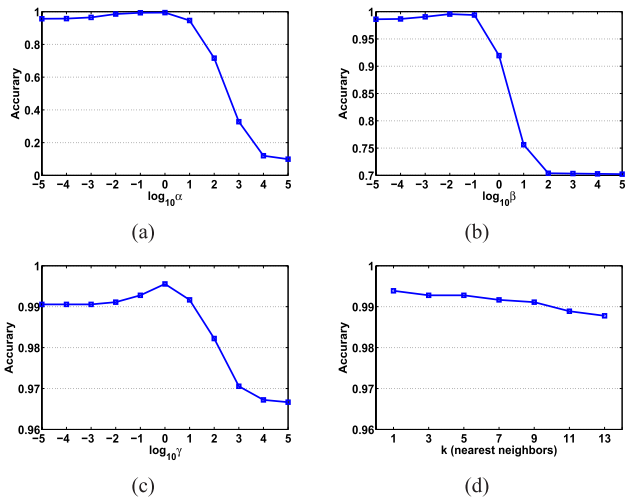

Fig. 9. Sensitivity analysis of the parameters: (a) varying α; (b) varying β; (c) varying γ; (d) varying k.

TABLE V
CLASSIFICATION ACCURACIES OF USING DIFFERENT KERNELS

	FAR=0.1%	FAR=1%	Rank-1
Linear	98.42%	99.66%	99.28%
RBF	73.42%	87.42%	89.28%
Polynomial	64.00%	91.56%	87.28%

same to the result (98.98%) when taking these 880 NIR images into transductive learning procedure. Note that, the corresponding results of second best approach SDA are 78.86% (out-of-sample) and 79.32% (transductive) respectively. In a word, our proposed THFM can also be generalized to out-of-sample patterns and achieve comparable results to the original transductive setting.

VII. CONCLUSION

This work has formulated the VIS-NIR face matching as a transductive learning problem. The modeling and the combination of good descriptor help improve the generalization performance of VIS-NIR matching problem. We particularly develop a novel transductive subspace learning method for heterogeneous face matching for domain invariant feature extraction. Our transductive model is discriminant and able to alleviate the cross-domain difference at the same time. To our best knowledge, it is the first attempt in this field to formulate a transductive learning for VIS-NIR matching. In addition, we give an in-depth analysis on the local feature based descriptors for VIS-NIR matching and and propose a Log-DoG based approach. We find that taking a feature normalization step would contribute a lot for extracting domain invariant to the final classification. Experiment results show that our proposed method has outperformed the related VIS-NIR matching approaches.

ACKNOWLEDGMENT

The authors would like to thank the editors and all anonymous reviewers for their constructive comments. They would also like to thank Z. Lei, who provided valuable feedback on this work.

REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[3] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Schölkopf, and A. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[4] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–7.

[5] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2707–2714.

[6] T. Chen, W. Yin, X. Zhou, D. Comaniciu, and T. Huang, "Total variation models for variable lighting face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1519–1524, Sep. 2006.

[7] J. Choi, S. Hu, S. S. Young, and L. S. Davis, "Thermal to visible face recognition," *Proc. SPIE*, vol. 8371, pp. 83711L-1–83711L-10, May 2012.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 886–893.

[11] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs NIR)," in *Proc. IEEE ICCVW*, Nov. 2011, pp. 2160–2167.

[12] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th ICML*, 1999, pp. 200–209.

[13] B. Klare and A. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Proc. 20th ICPR*, Aug. 2010, pp. 1513–1516.

[14] Z. Lei and S. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1123–1128.

[15] F. Li and H. Wechsler, "Open set face recognition using transduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1686–1697, Nov. 2005.

[16] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[17] S. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[18] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. 3rd Int. Conf. ICB*, Jun. 2009, pp. 209–218.

[19] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. 9th ECCV*, 2006, pp. 13–26.

[20] S. Liu, D. Yi, Z. Lei, and S. Li, "Heterogeneous face image matching using multi-scale features," in *Proc. 5th IAPRICB*, Mar./Apr. 2012, pp. 79–84.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[22] S. Pan, I. Tsang, J. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[23] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[24] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[25] H. Wang, S. Li, and Y. Wang, "Generalized quotient image," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun./Jul. 2004, pp. 498–505.

[26] R. Wang, J. Yang, D. Yi, and S. Li, "An analysis-by-synthesis method for heterogeneous face biometrics," in *Proc. 3rd Int. Conf. ICB*, 2009, pp. 319–326.

[27] Z. Xuan, K. Josef, and M. Kieron, "Ambient illumination variation removal by active near-ir imaging," in *Proc. Int. Conf. ICB*, Jan. 2006, pp. 19–25.

[28] X. Xie, J. Lai, and W. Zheng, "Extraction of illumination invariant facial features from a single image using nonsubsampled contourlet transform," *Pattern Recognit.*, vol. 43, no. 12, pp. 4177–4189, Dec. 2010.

[29] X. Xie, W.-S. Zheng, J. Lai, P. C. Yuen, and C. Y. Suen, "Normalization of face illumination based on large-and small-scale features," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1807–1821, Jul. 2011.

[30] D. Yi, S. Liao, Z. Lei, J. Sang, and S. Li, "Partial face matching between near infrared and visual images in MBGC portal challenge," in *Proc. 3rd Int. Conf. ICB*, Jun. 2009, pp. 733–742.

[31] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li, "Face matching between near infrared and visible light images," in *Proc. Int. Conf. ICB*, 2007, pp. 523–530.

[32] T. Zhang, Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, Nov. 2009.

[33] Z. Zhang, Y. Wang, and Z. Zhang, "Face synthesis from near-infrared to visual light via sparse representation," in *Proc. IEEE IJCB*, Oct. 2011, pp. 1–6.

Jun-Yong Zhu received the B.S. and M.S. degrees from the School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, China, in 2008 and 2010, respectively, where he is currently pursuing the Ph.D. degree with the Department of Applied Mathematics. His current research interests include heterogeneous face recognition, visual transfer learning using partial labeled or unlabeled auxiliary data, and nonlinear clustering. He has published several papers in international conferences, such as ICIP, AMFG, and ICDM. His cooperative ICDM 2010 paper won the Honorable Mention for Best Research Paper Awards and his CCBR 2012 paper won the Best Student Paper Awards.

Wei-Shi Zheng is currently an Associate Professor with Sun Yat-Sen University and a Key Member of Guangdong province introduced innovative computing science team. He joined the university under the one-hundred-people programme in 2011. His research direction is machine vision and intelligence learning. He is focusing on human-centered image understanding, including face recognition, person re-identification, and activity recognition. He has published over 40 major publications in top/leading journals (TPAMI, TNN, TIP, TSMC-B, PR) and top conferences (ICCV, CVPR, IJCAI). He has joined the organization of three tutorial presentations in ACCV 2012, ICPR 2012, and ICCV 2013 along with other colleagues in the CASIA. He has been awarded the New Star of Science and Technology of Guangzhou in 2012 and Guangdong Natural Science Funds for Distinguished Young Scholars in 2013.

Jian-Huang Lai received the M.Sc. degree in applied mathematics and the Ph.D. in mathematics from Sun Yat-Sen University, China, in 1989 and 1999, respectively. He joined Sun Yat-Sen University in 1989 as an Assistant Professor, where he is currently a Professor with the Department of Automation of School of Information Science and Technology and the Vice Dean of the School of Information Science and Technology. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet, and its applications. He has published over 80 scientific papers in international journals and conferences on image processing and pattern recognition, e.g., IEEE TPAMI, IEEE TNN, IEEE TIP, IEEE TSMC (Part B), Pattern Recognition, ICCV, CVPR, and ICDM. He serves as a Standing Member of the Image and Graphics Association of China and a Standing Director of the Image and Graphics Association of Guangdong.

Stan Z. Li received the M.Eng. degree from the National University of Defense Technology, China, and the Ph.D. degree from Surrey University, U.K. He is currently a Professor with the National Laboratory of Pattern Recognition and the Director of the Center for Biometrics and Security Research, Institute of Automation, and the Director of the Center for Visual Internet of Things Research, Chinese Academy of Sciences. He was with Microsoft Research Asia as a Researcher from 2000 to 2004. He was an Associate Professor with Nanyang Technological University, Singapore. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published over 200 papers in international journals and conferences, and authored and edited eight books. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and was an Editor-in-Chief for the *Encyclopedia of Biometrics*. He served as a Program Cochair for the International Conference on Biometrics in 2007 and 2009, a General Chair for the Ninth IEEE conference on Automatic Face and Gesture Recognition, and has been involved in organizing other international conferences and workshops.