Learning Mappings for Face Synthesis from Near Infrared to Visual Light Images

Jie Chen[1], Dong Yi[2], Jimei Yang[2,3], Guoying Zhao[1], Stan Z. Li[2], Matti Pietikäinen[1]

[1]Machine Vision Group, Department of Electrical and Information Engineering, University of Oulu, Finland
[2]Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing, 100190, China
[3]Department of Automation, University of Science and Technology of China
{jiechen, gyzhao, mkp}@ee.oulu.fi, {dyi, jmyang, szli}@cbsr.ia.ac.cn

Abstract

This paper deals with a new problem in face recognition research, in which the enrollment and query face samples are captured under different lighting conditions. In our case, the enrollment samples are visual light (VIS) images, whereas the query samples are taken under near infrared (NIR) condition. It is very difficult to directly match the face samples captured under these two lighting conditions due to their different visual appearances. In this paper, we propose a novel method for synthesizing VIS images from NIR images based on learning the mappings between images of different spectra (i.e., NIR and VIS). In our approach, we reduce the inter-spectral differences significantly, thus allowing effective matching between faces taken under different imaging conditions. Face recognition experiments clearly show the efficacy of the proposed approach.

1. Introduction

Face recognition has received increasing interest in recent years [19]. However, most current face recognition systems are designed for indoor, cooperative-user applications and the performance suffers from different environmental illumination. To this task, Li et al. present an active NIR imaging system (over a wavelength range of 0.7μm-1.1μm) for illumination invariant face recognition [10]. However, their system needs that both the enrollment and query samples are captured under NIR conditions. It is difficult for many applications since most face images are taken under visible light spectrum (over a wavelength range of 0.4μm-0.7μm), such as passport and driver license photos. Furthermore, some international organizations and standards recommend for taking VIS photos, e.g., the International Civil Aviation Organization, the International Organization for Standardization, and the International Electrotechnical Commission. In addition, most of the widely-used face datasets are provided under VIS conditions, e.g., FERET [15]. A straightforward matching between images of different spectra (i.e., NIR and VIS) is not effective mainly because of their different spectral properties (cf. Section 5.1). The proposed method is for those applications where it is administratively required to use VIS images on enrollment, such as E-passport and driver's license. To this end, we develop a transformation between the face images captured under NIR and VIS conditions. In this paper, we consider the synthesis of face images from NIR to VIS condition. Face synthesis from VIS to NIR can be accomplished similarly.

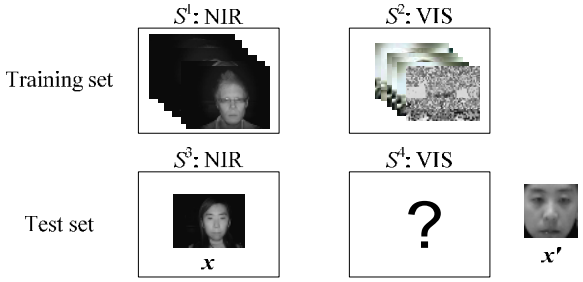

Fig. 1. Synthesize a virtual face sample x' in S^4 for an input face image x in S^3 given the training set S^1 and S^2.

Formally, as shown in Fig. 1, the problem to be discussed in this paper is that the enrollment face samples and query face samples are taken under different lighting conditions. In general, we have two training sets, S^1 (i.e., samples imaging under NIR condition) and S^2 (i.e., samples imaging under VIS condition). The elements in these two sets are the registered NIR-VIS pairs. Our goal is that for an input sample x under the NIR condition (i.e., $x \in S^3$ but $x \notin S^1$), we synthesize its corresponding virtual sample x' under the VIS condition (i.e., $x' \in S^4$). Here, the set S^3 is the test set whose samples are taken under the NIR condition and S^4 is the resulting set whose elements are virtual VIS images.

To perform the transformation from x to x', the proposed method includes two phases: training and testing procedures. In the training phase, we divide the samples in the training sets (S^1 and S^2) into patches. For each patch, we compute its feature. We then build up two patch dictionaries (D^t, t=1 and 2), and each element in D^t is a patch-feature pair. In the testing phase, for a testing sample x in the set S^3, we also perform the same procedure to obtain its patch-feature pairs. For each patch, we look up

the built dictionary $\boldsymbol{D}^1$ and compute its K nearest neighbors (KNN) in the same location. We then use the corresponding KNN patches in $\boldsymbol{D}^2$ to synthesize the face patch under the VIS condition. Combining each patch, a virtual sample $\boldsymbol{x}'$ in set S^4 is obtained. Finally we perform the recognition on the probing VIS images from the synthesized VIS ones. Face recognition experiments clearly show the efficacy of the proposed method.

The rest of this paper is organized as follows. Section 2 surveys some related works. Section 3, 4 and 5 describe the proposed framework and the facial feature used in this study, respectively. In Section 6, some experimental results are shown, followed by a conclusion in Section 7.

2. Related Work

To overcome the illumination variations in face domain, much effort has been made by modeling illumination on faces and correct illumination directions [19], such as [2]. The use of NIR imaging brings a new dimension [9, 10, 18]. A related work to ours is that of Yi et al. [18], who perform the matching from NIR faces to VIS ones directly using multi-variant regression by combining the linear discriminant analysis (LDA) and canonical correlation analysis (CCA). In [11], Lin and Tang proposed the common discriminant feature extraction, which can be seen as an extended version of linear discriminant analysis for two heterogeneous spaces. However, due to the fact that the illumination variations for the faces under NIR and VIS conditions are significantly different, the large intrapersonal differences decrease the matching performance seriously (cf. Section 5.1 and Section 6 for details). Therefore, we propose a type of mapping method for this problem, which significantly improves the matching performance between NIR and VIS images.

The mapping of one subject's sample from one lighting condition to another can be seen as a problem of image analogy (IA) [6]. Specifically, IA means that: Given a pair of images $\boldsymbol{y}$ and $\boldsymbol{y}'$ (the unfiltered and filtered source images, respectively), and another unfiltered target image $\boldsymbol{x}$, we need to find the mapping function that $\boldsymbol{y}'=\psi(\boldsymbol{y})$. Thus, for a new input image $\boldsymbol{x}$, we have $\boldsymbol{x}'=\psi(\boldsymbol{x})$. An interesting work is the face sketch synthesis [12, 17]. Meanwhile, Tang and Wang [17] apply principal component analysis (PCA) for this task. Liu et al. [12] use a locally linear mapping and use Euclidean distance as a measure between face patches. However, because the Euclidean distance used in [12] is not an illumination-invariant feature, the assumption of a local geometry preserving in [12] is not suitable for our problem (lighting changes significantly from NIR to VIS condition). Neither is PCA. Different from their methods [12, 17], we use local binary pattern (LBP) which is an illumination-invariant feature [1, 10, 14] and validates the assumption of a local geometry preserving (cf. Section 5.2 for more details).

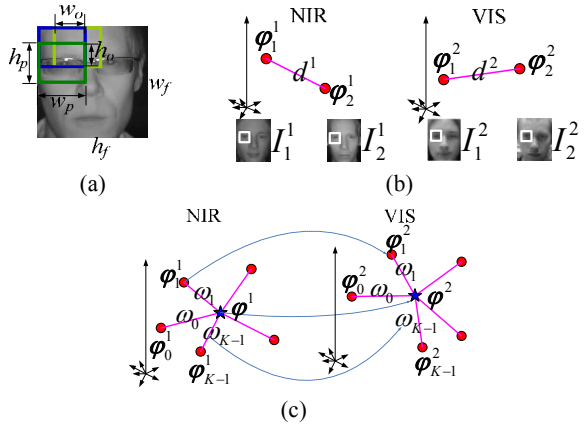

Fig. 2. Local linear mapping between patches: (a) a face sample (size $w_f \times h_f$) is divided into patches (size $w_p \times h_p$) in an overlapped way (size $w_o \times h_o$); (b) distances between two patch pairs in different image spaces (i.e., NIR and VIS, respectively) and (c) a patch and its K-nearest neighbors in different manifold spaces.

3. Face Synthesis by Learning

3.1. Modeling of NIR and VIS images

First we formulate the relationship between a NIR face image and its counterpart under VIS condition based on manifold [16]. We show that there exists a local linear mapping function $\psi(\cdot)$ between the NIR and VIS samples by assuming that both the NIR and VIS images are respectively captured under similar lighting conditions and nearly homogenous illumination on face. Thus, a regression-based solution is applicable to a novel image.

More specifically, as shown in Fig. 2 (a), we divide the samples in the sets of S^1 and S^2 into patches in an overlapped way as [1]. For two subjects (i.e., O_1 and O_2,), as shown in Fig. 2 (b), we use the superscript to denote the image set and the subscript as the index of images. For example, the superscript "t" of I_1^t and I_2^t (t=1,2) denotes that both are from the set S^t. However, the two pairs (I_1^1, I_1^2) and (I_2^1, I_2^2) are the photos of the same subject (i.e., O_1 and O_2,), respectively. $\boldsymbol{\varphi}_i^t$ (t=1, 2 and i=1, 2) are the patches from I_i^t in the same location as shown in Fig. 2 (b).

We have the following statement:

$$KNN(\boldsymbol{\varphi}_1^1, \boldsymbol{\varphi}_2^1) \Rightarrow KNN(\boldsymbol{\varphi}_1^2, \boldsymbol{\varphi}_2^2) \text{ and } d^1/d^2 \approx \xi. \quad (1)$$

Here, $KNN(x,y)$ denotes that x and y are of K nearest neighbors; $d^t=\| \boldsymbol{\varphi}_1^t - \boldsymbol{\varphi}_2^t \|$ (t=1, 2), i.e., the distance between $\boldsymbol{\varphi}_1^t$ and $\boldsymbol{\varphi}_2^t$. The statement in Eq. (1) says there exits a local geometry preserving. In other words, if $\boldsymbol{\varphi}_1^1$ and $\boldsymbol{\varphi}_2^1$ are of K nearest neighbors, so are $\boldsymbol{\varphi}_1^2$ and $\boldsymbol{\varphi}_2^2$; and the distance ratio d^1/d^2 approaches a constant ξ.

Roweis and Saul develop the locally linear embedding (LLE) method, recovering global nonlinear structure from locally linear fits [16]. Following their idea, if we learn the

manifold of the patches $\boldsymbol{\varphi}_i^1$ (i=0,1,…, N-1, N=|S^1|, i.e., the cardinality of S^1), which are from the same location of different subjects but under NIR lighting, we obtain a NIR manifold $\mathcal{M}^1$. Likewise, for the patches $\boldsymbol{\varphi}_i^2$ (i=0, 1, …, N-1, N=|S^2|) we learn a VIS manifold $\mathcal{M}^2$. From the statement in Eq. (1), we have that $\mathcal{M}^1$ and $\mathcal{M}^2$ are approximately isometric.

Thus, as shown in Fig. 2 (c), for one patch $\boldsymbol{\varphi}^1$ in $\mathcal{M}^1$, there is a locally linear fit between $\boldsymbol{\varphi}^1$ and its K nearest neighbors $\boldsymbol{\varphi}_k^1$ (k=0, 1, …, K-1) [16] and

$$\boldsymbol{\varphi}^1 = \psi^1(\omega_k, \boldsymbol{\varphi}_k^1) = \sum_{k=0}^{K-1} \omega_k \boldsymbol{\varphi}_k^1, \tag{2}$$

where ω_k is a group of weights computed based on the distances between them. According to the local geometry preserving between $\mathcal{M}^1$ and $\mathcal{M}^2$, we map the patches $\boldsymbol{\varphi}^1$ and $\boldsymbol{\varphi}_k^1$ from $\mathcal{M}^1$ to $\mathcal{M}^2$ and get their counterparts in $\mathcal{M}^2$ (i.e., $\boldsymbol{\varphi}^2$ and its K nearest neighbors $\boldsymbol{\varphi}_k^2$). In addition, we also borrow the weights from $\mathcal{M}^1$ since $\mathcal{M}^1$ and $\mathcal{M}^2$ are approximately isometric. Correspondingly, the locally linear fit between $\boldsymbol{\varphi}^2$ and $\boldsymbol{\varphi}_k^2$ is approximately as follows:

$$\boldsymbol{\varphi}^2 = \psi^2(\omega_k, \boldsymbol{\varphi}_k^2) = \sum_{k=0}^{K-1} \omega_k \boldsymbol{\varphi}_k^2, \tag{3}$$

Note that $\boldsymbol{\varphi}^1$ and $\boldsymbol{\varphi}^2$ are two patches from the same location of the same subject but under different lighting conditions (i.e., NIR and VIS). Likewise, $\boldsymbol{\varphi}_k^1$ and $\boldsymbol{\varphi}_k^2$ are patches from the same locations of the same subjects and under different lightings, respectively.

As mentioned above, we call the mapping from NIR images to VIS ones as a local linear mapping since it combines the two factors: firstly, the *locally linear* fit between a patch $\boldsymbol{\varphi}^t$ and its K nearest neighbors $\boldsymbol{\varphi}_k^t$ (t=1, 2) as shown in Eqs. (2) and (3); secondly, the *mapping* of the patches $\boldsymbol{\varphi}^1$ and $\boldsymbol{\varphi}_k^1$ from $\mathcal{M}^1$ to their counterparts $\boldsymbol{\varphi}^2$ and $\boldsymbol{\varphi}_k^2$ in $\mathcal{M}^2$. Thus, for each patch $\boldsymbol{\varphi}_i^1$ of an input image $\boldsymbol{x}$ under NIR condition, we can use the locally linear mapping to synthesize its counterpart $\boldsymbol{\varphi}_i^2$ under VIS condition given the patch pairs ($\boldsymbol{\varphi}_{i,k}^1$, $\boldsymbol{\varphi}_{i,k}^2$) ($k$=0,1,…,$K$-1). Combining each virtual patch of the image $\boldsymbol{x}$ under NIR condition, we synthesize its counterpart $\boldsymbol{x}'$ under VIS condition.

As demonstrated in [16], the two locally linear fits $\psi^1(\bullet,\bullet)$ and $\psi^2(\bullet,\bullet)$ for VIS and NIR manifolds in Eq. (2) and (3) both take the assumption that the data points used for the manifold learning should be well-sampled. In our case, dividing face samples into patches reduces the dimensionality of data points, which in turn reduces the demand on the number of samples. Furthermore, neurophysiological evidence hints us that we can find the similar patches for an input patch. Specifically, one can imagine the patches (e.g., eyes, nose, and mouth) being analyzed in parallel. This might be an explanation that we say one person having another one's eyes [7]. The procedure discussed in this section shows such a parallel approach naturally.

Note that in our implementation we just borrow the idea of manifold learning of [16] to make the statement in Eq. (1) to learn an implicit local linear mapping between the patches of NIR and VIS conditions but we in fact do not learn an explicit global manifold.

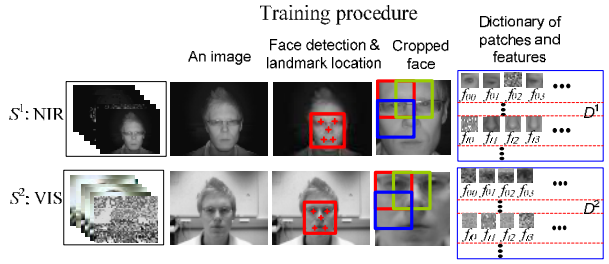

Fig. 3. Training procedure.

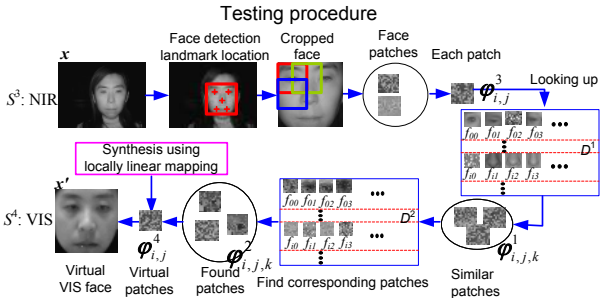

Fig. 4. Testing procedure.

3.2. Training phase

As shown in Fig. 3, during the training procedure, we automatically detect the present face by a face detector and locate its landmarks for each image in the training sets, S^1 and S^2. Here, the landmarks consist of 5 points, i.e., two eyes, nose and two mouth corners. Using the landmarks, each detected face is cropped and normalized to the size $w_f \times h_f$ and then divided into patches (size $w_p \times h_p$) in an overlapped way (size $w_o \times h_o$) as shown in Fig. 2 (a). Let $\boldsymbol{\varphi}_{i,j}$ denote a patch, and i=0, 1, …, m-1; j=0, 1, …, n-1, where m and n are the number of patches of a sample in the row and column direction, respectively. Specifically,

$$m=(h_f-h_p)/(h_p-h_o)+1 \text{ and } n=(w_f-w_p)/(w_p-w_o)+1. \tag{4}$$

For each patch $\boldsymbol{\varphi}_{i,j}$, we compute its feature $\boldsymbol{f}_{i,j}$ (cf. Section 4). Combining face patches $\boldsymbol{\varphi}_{i,j}$ and their features $\boldsymbol{f}_{i,j}$ of all faces in the two training sets, we obtain two dictionaries $\boldsymbol{D}^t=\{(\boldsymbol{\varphi}_{i,j}, \boldsymbol{f}_{i,j})\}^t$ (t=1, 2), one for S^1 (i.e., NIR) and the other for S^2 (i.e., VIS). Here, the dimensionality of $\boldsymbol{D}^t$ is of $M \times N$, where M= $m \times n$, and N=|S^1|=|S^2|. More specifically, each row of $\boldsymbol{D}^t$ is composed of M patch-feature pairs, each one coming from the same location in the face samples.

3.3. Testing phase

During the testing procedure, as shown in Fig. 4, for an input image $\boldsymbol{x}$ in S^3, we also perform the same procedure as in the training phase and obtain the patch-feature pairs $(\boldsymbol{\varphi}_{i,j}^3, f_{i,j}^3)$. For each patch $\boldsymbol{\varphi}_{i,j}^3$, as discussed in Section 3.1, we look up the dictionary $\boldsymbol{D}^1$ for its K nearest neighbors $\boldsymbol{\varphi}_{i,j,k}^1$ (k=0,1,…,K-1 and the superscript "1" denotes from $\boldsymbol{D}^1$) by its feature $f_{i,j}^3$. Note that for each input patch $\boldsymbol{\varphi}_{i,j}^3$, we only look up those patches with the same location as $\boldsymbol{\varphi}_{i,j}^3$ i.e., only the row ($r = i\times n+j$) in $\boldsymbol{D}^1$.

Here, we need to compute the distance between two given feature vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. Because we employ the LBP histogram to represent each patch (cf. Section 4), we use the histogram intersection Ψ (H_1,H_2) as a similarity measure of two histograms H_1 and H_2:

$$\gamma = \Psi(H_1, H_2) = \sum_{i=1}^{L} \min(H_{1,i}, H_{2,i}), \tag{5}$$

where L is the number of bins of a histogram and γ represents the computed similarity measure for simplicity. Note that both H_1 and H_2 are normalized beforehand.

For each of the K nearest neighbors $\boldsymbol{\varphi}_{i,j,k}^1$, we find its corresponding patch $\boldsymbol{\varphi}_{i,j,k}^2$ in dictionary $\boldsymbol{D}^2$. We then synthesize the virtual patch $\boldsymbol{\varphi}_{i,j}^4$ (under VIS condition) of current one $\boldsymbol{\varphi}_{i,j}^3$ (a NIR one) using the locally linear mapping (cf. Section 3.1) with its neighbors $\boldsymbol{\varphi}_{i,j,k}^2$:

$$\boldsymbol{\varphi}_{i,j}^4 = \sum_{k=0}^{K-1} \omega_k \boldsymbol{\varphi}_{i,j,k}^2, \tag{6}$$

where ω_k are the normalized weights:

$$\omega_k = \gamma_k / \sum_{i=0}^{K-1} \gamma_i, \tag{7}$$

and γ_k (k=0, 1, …, K-1) are the similarities between the input patch $\boldsymbol{\varphi}_{i,j}^3$ and its K nearest neighbors $\boldsymbol{\varphi}_{i,j,k}^1$. Combining each generated patch $\boldsymbol{\varphi}_{i,j}^4$, a virtual face sample $\boldsymbol{x}'$ in S^4 is obtained. Note that for a pixel in the overlapped region between patches, its value is the average of those patches that cover that pixel.

4. Facial Features

4.1. Local binary pattern

In our case, we use LBP to represent a face patch due to its discriminative power and computational efficiency [14]. The basic form of LBP is illustrated in Fig. 5 (a) and (b). The operator takes as input a local neighborhood around each pixel and thresholds the neighborhood pixels at the value of the central pixel. The resulting binary-valued string is then weighted as follows:

$$LBP(I_c) = \sum_{j=0}^{p-1} 2^p s(I_j - I_c), \tag{8}$$

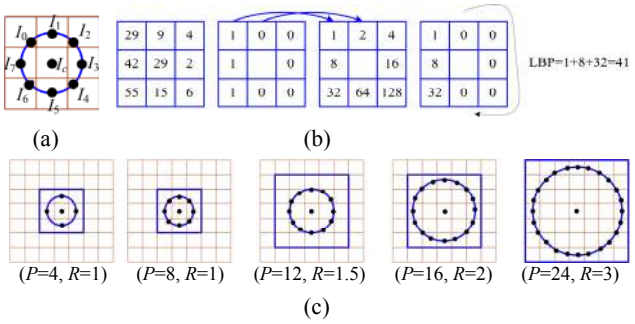

Fig. 5. LBP: (a) a pixel and its eight neighbors; (b) the basic LBP; and (c) circularly symmetric neighbor sets for different (P, R).

where p runs over the neighbors. I_c and I_j are the gray-level values at c and j, and $s(A)$ is 1 if $A \geq 0$ and 0 otherwise.

Two extensions of the original LBP are made in [14]. One is the multi-resolution operator as depicted in Fig. 5 (c). Here, we use $\text{LBP}_{\text{P,R}}$ to denote its parameters (e.g., $\text{LBP}_{4,1}$). The other is the *uniform patterns*: an LBP is 'uniform' if it contains at most one 0-1 or 1-0 transition when viewed it as a circular bit string. Another extension is the center-symmetric LBP (CSLBP) [5].

For each patch $\boldsymbol{\varphi}_{i,j}$ of a sample from S^1 or S^2, we compute its LBP histogram feature $\boldsymbol{f}_{i,j}$, which might be computed in different resolution as shown in Fig. 5 (c). We will discuss it in the following subsection.

4.2. Multi-resolution LBP

We employ the multi-resolution LBP for our task since it improves the performance of a single resolution of LBP significantly [1, 14]. Inspired by the idea in [8], we combine the outputs of multi-resolution LBP. We call this *MLBP*, where "*M*" is an abbreviation of multi-resolution. However, in our implementation we do not combine the classifier for a classification task but for the combination of a set of distance measures between patches. Specifically, for a patch $\boldsymbol{\varphi}_{i,j}$, its feature $\boldsymbol{f}_{i,j}$ ={{$H_{i,j}$(CSLBP$_{8,1}$)}, {$H_{i,j}$(LBP$_{\text{P,R}}$)}}, where (P, R)=(4, 1), (8, 1), (12, 1.5), (8, 2), (16, 2), (16, 3) and (24, 3). In other words, $\boldsymbol{f}_{i,j}$ is composed of C=8 histogram components. These C components are not concatenated to one histogram as in [1, 14] but each component is stored in an array, respectively. Here, we use "{}" to contain each component to denote their independence. In this case, there are C distance components between the input patch $\boldsymbol{\varphi}_{i,j}$ and one of its neighbors $\boldsymbol{\varphi}_{i,j,k}$ since $\boldsymbol{f}_{i,j}$ is composed of C histogram components. We denote these C distance measures as $\boldsymbol{\gamma} = \{\gamma_i, i = 0, 1, ..., C-1\}$.

The algorithm of using MLBP to synthesize a virtual VIS sample is as follows:

- Compute each histogram component of MLBP over each patch of each sample in the sets S^1 and S^2, i.e., $\{H_{s,q,c}^{Tr}; s=0,1,\ldots, N\text{-}1; q=0,1,\ldots, M\text{-}1, c=0,1,\ldots,C\text{-}1\}$. Here, "*Tr*" denotes training set; N=|S^1|=|S^2|; M is the

number of the patches of a sample; C is the number of the components of MLBP.

- Compute each histogram component of MLBP over each patch of an input face sample, i.e., $\{H_{q,c}^{In}$, q=0, 1, …, M-1, c=0, 1, …, C-1$\}$. Here, "*In*" denotes the input sample.
- Compute the similarities between each patch of the input sample and the patches in the same location of samples in the training set S^1 using Eq.(5): $\Psi(H_1,H_2)=\sum_{i=1}^{L}\min(H_{1,i},H_{2,i})$, here $H_1=H_{q,c}^{In}$ and $H_2=H_{s,q,c}^{Tr}$. We denote these similarities as $\{\gamma_{q,s,c}$, q=0, 1, …, N-1; s=0, 1, …, M-1, c=0, 1, …, C-1$\}$. In other words, for s^{th} patch of the input sample, we have C similarities with the s^{th} patch of each sample in the training set S^1.
- Combine the similarities: $\gamma_{q,s}=f(\gamma_{q,s,c})$. Here, f is a function of combining these C similarities. In our case, we experientially use the product rule [8]:

$$\gamma_{q,s}=f(\gamma_{q,s,c})=\prod_{c=0}^{C-1}\gamma_{q,s,c}. \quad (9)$$

- Find those K nearest neighbors for each patch of the input sample using the combined similarities $\gamma_{q,s}$, compute the corresponding weights and synthesize the virtual sample as shown in Section 3.3.

5. Validation of the Proposed Method

In the section, we validate the two facts: one is that it is difficult to perform the direct match between images of different spectra (i.e., NIR and VIS); the other is the statement in Eq. (1) in Section 3.1.

5.1. Match between NIR and VIS images

Motivated by [4], we adopt a Lambertian model of image formation. An image $I(x,y)$ under a point light source is expressed as:

$$I(x,y)=\boldsymbol{n}(x,y)\boldsymbol{s}\int E(\lambda,x,y)S(\lambda,x,y)Q(\lambda)d\lambda, \quad (10)$$

where $E(\lambda,x,y)$ is a spectral power distribution (SPD) of an incident light; $S(\lambda,x,y)$ is the surface reflectance function; $Q(\lambda)$ denotes the spectral sensitivity of the camera sensor; $\boldsymbol{n}(x,y)$ is the surface normal in the 3D space; $\boldsymbol{s}$ is the lighting direction (a column vector, with magnitude). We assume that the camera behaves as an exact Dirac delta function [4]. Thus, we have $Q(\lambda)=q\delta(\lambda)$ and Eq. (10) is written as:

$$I(x,y)=\boldsymbol{ns}E(\lambda)S(\lambda)q. \quad (11)$$

As we use LBP histogram to represent the images, and each LBP string is composed of several bits [14]. Each bit is obtained by thresholding its neighbors I_j using the central pixel I_c as shown in Eq. (8). Equivalently, Eq. (8) can be rewritten as:

$$LBP(I_c)=\sum_{j=0}^{p-1}2^p\mathbf{1}(I_j/I_c), \quad (12)$$

where $\mathbf{1}(A)$ is 1 if $A\geq 1$ and 0 otherwise.

According to Eq. (11), for each bit of a LBP string we have:

$$\frac{I_j}{I_c}=\frac{[\boldsymbol{ns}E(\lambda)S(\lambda)q]_j}{[\boldsymbol{ns}E(\lambda)S(\lambda)q]_c}. \quad (13)$$

Both the NIR and VIS images are respectively captured under the similar lighting conditions and nearly homogenous illumination on face. Thus, we can assume that the lighting direction $\boldsymbol{s}$ between the neighboring pixels is similar. According to Eq. (13), for each bit we have:

$$b_j=\mathbf{1}\left(\frac{I_j}{I_c}\right)\approx\mathbf{1}\left(\frac{[\boldsymbol{n}S(\lambda)]_j}{[\boldsymbol{n}S(\lambda)]_c}\right). \quad (14)$$

Here, we denote the surface reflectance ratio as

$$R_{sr}=[\boldsymbol{n}S(\lambda)]_j/[\boldsymbol{n}S(\lambda)]_c. \quad (15)$$

Thus, we have

$$b_j=\mathbf{1}(R_{sr}). \quad (16)$$

For those pixels from a patch $\boldsymbol{\varphi}_i^t$ (t=1, 2 and i=1, 2) as shown in Fig. 2 (b), we denote a bit of a LBP string of a pixel as $b_j(\boldsymbol{\varphi}_i^t)$. We have the statement that the difference between $b_j(\boldsymbol{\varphi}_1^1)$ and $b_j(\boldsymbol{\varphi}_1^2)$ is nonlinear. Specifically, the wavelength from NIR (i.e., λ_{NIR} for $\boldsymbol{\varphi}_1^1$) to VIS (i.e., λ_{VIS} for $\boldsymbol{\varphi}_1^2$) changes significantly. In addition, the surface reflectance $S(\lambda)$ of human skin is a nonlinear function of wavelength λ [3] and surface normal $\boldsymbol{n}$ could also vary between pixels. Thus, according to Eq. (15), the surface reflectance ratio R_{sr} of neighbors also varies nonlinearly from λ_{NIR} to λ_{VIS}. According to Eq. (16), it leads to the nonlinear difference between $b_j(\boldsymbol{\varphi}_1^1)$ and $b_j(\boldsymbol{\varphi}_1^2)$, which nonlinearly changes the LBP string of the current pixel and so the LBP histogram of a patch. Therefore, it is difficult to perform a direct match between VIS and NIR samples.

5.2. Validation of local geometry preserving

It is difficult to mathematically prove the statement in Eq.(1) in Section 3.1 because it is difficult to model surface reflectance $S(\lambda)$ of human skin due to its complicated histological tissue [3, 13]. However, we attempt to state Eq.(1) by the aid of statistical evidences. More specifically, for each patch $\boldsymbol{\varphi}_i^1$ in dictionary $\boldsymbol{D}^1$, we compute its K nearest neighbors $\boldsymbol{\varphi}_{i,k}^1$ in $\boldsymbol{D}^1$ and the corresponding K distances $d_{i,k}^1$ (k=0,1,…,K-1) between $\boldsymbol{\varphi}_i^1$ and its neighbors (cf. Fig.2 (c)). Likewise, for each patch $\boldsymbol{\varphi}_j^2$ in $\boldsymbol{D}^2$, we also compute its K nearest neighbors $\boldsymbol{\varphi}_{j,k}^2$ in $\boldsymbol{D}^2$ and the corresponding K distances $d_{j,k}^2$.

To validate the statement in Eq. (1), we should validate

the following two propositions: Proposition-1 is that if ($\boldsymbol{\varphi}_i^1$, $\boldsymbol{\varphi}_j^2$) is a registered NIR-VIS pair, the K pairs ($\boldsymbol{\varphi}_{i,k}^1$, $\boldsymbol{\varphi}_{j,k}^2$) are also registered NIR-VIS pairs, respectively. Proposition-2 is that $d_{i,k}^1$ varies directly as $d_{j,k}^2$.

Table 1. Comparison of correctly matching KNN and distance

	Euclidean	PCA	Our method
Ratio(*KNN*)	80.4	87.7	92.5
average of distance ratio and standard variance	1.000(0.001)	0.998(0.001)	1.003(0.008)

As shown in Table 1, we give the statistical evidences on 1,500 images and each image is divided into patches as shown in Fig. 2 (a). The first row in this table is the percentage of the correctly matching KNN between ($\boldsymbol{\varphi}_{i,k}^1$, $\boldsymbol{\varphi}_{j,k}^2$), which is computed as:

$$Ratio(KNN)= \frac{\#correctly\ matching}{\#total\ patches} . \quad (17)$$

Here, a correctly matching is that both ($\boldsymbol{\varphi}_i^1$, $\boldsymbol{\varphi}_j^2$) and its K pairs ($\boldsymbol{\varphi}_{i,k}^1$, $\boldsymbol{\varphi}_{j,k}^2$) are registered, respectively. The second row is the average of the distance ratios between $d_{i,k}^1$ and $d_{j,k}^2$ which are correctly matched and the standard variances in parentheses.

Likewise, we also show the values using the Euclidean distance as in [12] and PCA as a measure to compute the distances between patches and verify whether the Euclidean distance and PCA can satisfy the two propositions. From the table, one can find that Proposition-1 is approached quite well by our method. It performs much better than PCA and Euclidean distance. In addition, all of the three methods can match Proposition-2 quite well.

6. Experimental Results

In this section, we present two groups of experiments with the proposed methods and compare them to some existing algorithms. For the first group of experiments, all the images in S^1 and S^2 are taken under similar and *homogenous* lighting conditions, respectively. Specifically, the samples in S^1 are captured using the active NIR conditions as in [10] and the samples in S^2 are captured under visual lighting indoors. Some examples are shown in Fig. 6. All of them are transformed into eight bit intensity images.

For the second group of experiments, all the images in S^1 and S^2 are taken under *heterogeneous* lighting conditions, respectively. Specifically, the samples in S^2 are captured indoors under three different visual lighting: normal, weak and dark. Normal illumination means that good environmental lighting is used. Weak illumination

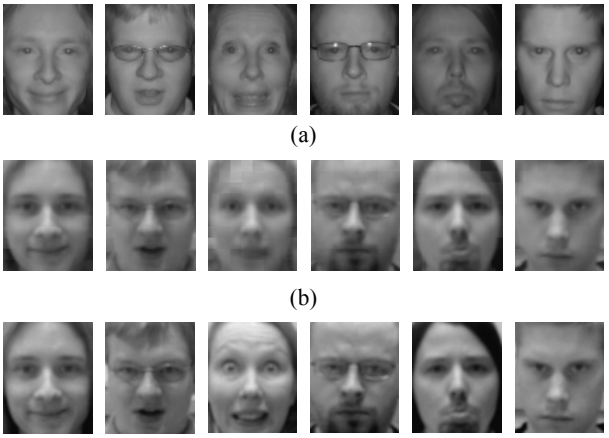

Fig. 6. Synthesized images; (a) input images under NIR condition; (b) synthesized images using MLBP; (c) Ground truth under VIS condition.

Fig. 7. Faces captured under different illuminations; (a) images under NIR condition with the three environmental illuminations: normal, weak and dark (from left to right); (b) images under three different VIS conditions: normal, weak and dark (from left to right).

means that only computer display is on and each subject sits on the chair in front of the computer. Dark illumination means near darkness with computer display being shut down. The intensity of light (lux) values for these three cases are 800, 50 and near 0 lx, respectively. Likewise, the samples in S^1 are captured using the active NIR imaging system as in [10] with the same three environmental illuminations. Some examples are shown in Fig. 7. From this figure, one can find that the illuminations in S^2 vary significantly. However, using the active NIR imaging system, the unfavorable lighting is almost unseen in the NIR face images in S^1.

6.1. Experiments on homogenous illumination

In this section, we will present the first group of experiments.

6.1.1. Setup

In our implementation, both S^1 and S^2 are composed of N=1,500 samples, which include 250 subjects and each subject containing 6 images (But they are in different expressions, such as happiness and disgust.).

For each normalized sample, as shown in Fig. 2 (a), we experientially set the value of each parameter as follows: w_f=64, h_f=80; w_p=16, h_p=16; and w_o =12, h_o =12. Thus, m=17 and n=13. During testing, we use the leave-one-out

strategy and set K=15 based on experiments. In addition, we use uniform patterns for different resolution LBPs. For how to set the parameter values (i.e., (w_f, h_f), (w_p, h_p), (w_o, h_o)), please refer to [1].

6.1.2. Synthesized samples

Some virtual samples are shown in Fig. 6 (b). From this figure, one can find that the synthesized images look fine subjectively except the third one. We believe that her extreme expression makes it difficult to find the similar patches in the training set. Collecting more samples in the training set is helpful (as discussed in Section 3.1).

6.1.3. Face recognition performance

In this section, we report the face recognition accuracies using some existing methods on our task, i.e., the methods of [6], [17], [12] and [18]. As shown in Table 2, "No" denotes that we directly match the face samples; "IA" is the method of Hertzmann et al. [6]; "PCA" is that of Tang and Wang [17]; "E_LLM" (an abbreviation for Euclidean plus local linear mapping) is that of Liu et al. [12]; and "LDA+CCA" is the method of [18]. Note that all of these methods were implemented by us following their ideas.

The data sets used in this part include: 1) the set S^1 as shown in Fig. 1, in which all images are captured under NIR condition; 2) the set S^2, in which all images are captured under VIS condition; 3) the sets $S^{4,i}$ (i=1, 2, 3, 4), in which all images are synthesized by IA, PCA, E_LLM and our method, respectively. All of these four sets are composed of 250 subjects, each showing 6 samples.

As shown in Table 2, all of cases use the same training set S^1 but different testing sets. Specifically, "No" denotes that we use the set S^2 as the testing set directly, which is a baseline. "IA" means that we use the virtual set $S^{4,1}$ as the testing set. Likewise, "PCA" and "E_LLM" means that we use the sets $S^{4,2}$ and $S^{4,3}$ as the testing sets. For our method, we use the set $S^{4,4}$. In addition, for "LDA+CCA", we use the set S^2 directly.

As the feature we use the concatenated multi-resolution LBP histogram as in [1]. We divide samples into patches as shown in Fig. 2 (a). For any two images, we compute the histogram intersection of the concatenated LBP features of each patch. We then sum the similarities of all the patches as the measure between two images. The nearest neighbor method is used as a classifier. We compare each sample in the testing sets with all samples in the training set.

The cumulative match scores for the six methods are shown in Table 2. The results clearly demonstrate the good performance of our algorithm. Using S^2 as a testing set, the first match for "No" is only 2% and the tenth rank is no more than 21%. It shows that it is difficult to directly match the images under VIS and NIR conditions since they are captured under completely different lightings (cf. Section 5.1). However, all synthesized sets $S^{4,i}$ (i=1, 2, 3, 4) improve the recognition performance significantly (e.g., 76.8%, 85.0%, 87.0% and 94.2% at rank 1, respectively).

Table 2. Cumulative match scores for the six methods on homogeneous illumination (%)

Rank	1	3	5	7	10	15	20
No	2.0	13.0	15.3	18.3	20.7	27.0	31.0
IA	76.8	81.3	85.3	88.7	90.0	94.3	95.2
PCA	85.0	87.2	89.5	92.3	94.7	95.0	97.3
E_LLM	87.0	89.3	90.4	93.0	93.0	93.5	93.5
LDA+CCA	96.5	97.1	98.3	98.6	98.7	98.8	99.1
Ours	94.2	95.1	96.6	96.9	98.4	99.2	100

Table 3. Cumulative match scores for the six methods on heterogeneous illumination (%)

Rank	1	3	5	7	10	15	20
No	3.0	9.0	13.0	18.3	24.0	24.0	24.0
IA	86.3	87.7	90.0	91.7	92.7	95.3	96.0
PCA	92.0	93.3	94.7	95.0	95.7	96.0	96.3
E_LLM	93.3	94.3	95.0	95.0	95.7	96.0	96.0
LDA+CCA	82.0	83.0	85.0	88.0	90.0	93.0	96.0
Ours	97.3	97.7	98.3	98.3	98.7	98.7	99.3

In comparison to the method of "LDA+CCA" [18], our method is more suitable for a large scale dataset and is very efficient for incremental learning. Specifically, Yi et al. use LDA and CCA for the NIR-VIS face image matching problem. Their method needs to compute the project matrix for LDA and this matrix needs to be re-computed once a new sample is added. Thus, adding new samples to a large scale dataset decreases significantly the efficiency of their method. On the other hand, our method computes more efficiently by using the LBP descriptor. It computes the features for each sample separately without any project matrix, which is quite suitable for a large scale dataset. Furthermore, our method can save the feature of each sample for future use. When new samples are added, we just need to compute the features for the new one. This makes our approach very efficient for incremental learning.

6.2. Experiments on heterogeneous illumination

In this section, we will present the second group of experiments. In our implementation, both S^1 and S^2 include 50 subjects. Specifically, S^1 is composed of N=300 samples, 6 images for each subject (2 samples for each of

the three different illuminations). Now S^2 is composed of N=100 samples and 2 images for each subject but only in normal illumination. In the set S^2, we do not include the samples taken in weak and dark illuminations since we hope to synthesize those virtual face samples in normal illumination for face recognition with a good performance. During testing, we use the same setup as in Section 6.1.

We also compare the performance of our methods with those of [6], [17], [12] and [18]. The cumulative match scores for the six methods are shown in Table 3. The results clearly demonstrate the superiority of our algorithm even the set S^1 is taken under heterogeneous lighting conditions. It shows that the assumption in Section 3.1 (i.e., both the NIR and VIS images should be captured under similar lighting conditions and nearly homogenous illumination, respectively) is not strictly required for an active NIR imaging system. In addition, the higher performance of these methods in Table 3 compared to that of in Table 2 (expect the method "LDA+CCA" of [18]) is twofold: one is that the unfavorable illumination variations are almost unseen in the NIR face images in S^1; and the other is that both the number of subjects and the size of the test sets are smaller than those in Table 2.

The performance of the method "LDA+CCA" decreases significantly. It is because that this method needs a pairwise training set, and so we use 100 samples of S^1 (only under normal illumination) and all S^2 for training, and we test the method by the other samples in S^1 versus S^2. Due to the small training set (100 images and 2 images per subject), the performance of this method decreases significantly (cf. Table 2 and Table 3). In this case, learning a mapping for face synthesis demonstrates its superiority as shown in Table 3.

7. Conclusion

In this paper we focus on a new problem in which the enrollment and query face samples are captured under different lighting conditions. To this end, we proposed a patch based transformation method with which a virtual sample is synthesized from an input sample. By this way, we reduce the intrapersonal difference caused by the completely different lightings (VIS vs. NIR). Experimental results show that the synthesized samples by the proposed method improves face recognition performance significantly (e.g., from 2.0% to 94.2% at rank 1 under homogenous illumination and from 3.0% to 97.3% at rank 1 under heterogeneous illuminations). Future work will focus on the problem that how to use those existing face datasets under VIS conditions but without the NIR pairs to facilitate training and testing procedure.

Acknowledgements

We would like to thank Timo Ahonen, who suggested improvements on the validation of the proposed method. This work was supported by the Academy of Finland, the Finnish Funding Agency for Technology and Innovation, the European Regional Development Fund, the Chinese National Natural Science Foundation Project #60518002, and the Chinese National Hi-Tech (863) Program Project #2008AA01Z124.

Reference

[1] T. Ahonen, A. Hadid and M. Pietikäinen, Face description with local binary patterns: application to face recognition, *PAMI*, 2006.

[2] R. Basri and D.W. Jacobs, Lambertian reflectance and linear subspaces, *PAMI* 2003.

[3] R. Doornbos, R. Lang, M. Aalders, F. Cross and H. Sterenborg, The determination of in vivo human tissue optical properties and absolute chromophore concentrations using spatially resolved steady-state diffuse reflectance spectroscopy, *Physics in Medicine and Biology.* 1999

[4] G. D. Finlayson, S.D. Hordley, C Lu, M.S. Drew, On the removal of shadows from images, *PAMI*, 2006

[5] M. Heikkilä, M. Pietikäinen and C. Schmid. Description of interest regions with local binary patterns. *PR,* 2008.

[6] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, D. Salesin, Image analogies, *SIGGRAPH*, 2001.

[7] M. Kirby and L. Sirovich, Application of the karhunen-loève procedure for the characterization of human faces, *PAMI*, 1990.

[8] J. Kittler, M. Hatef, R. P.W. Duin and J. Matas, On combining classifiers, *PAMI*, 1998.

[9] Z. Lei, Q. Bai, R. He, S. Z. Li. Face shape recovery from a single image using CCA mapping between tensor spaces. *CVPR*, 2008.

[10] S. Z. Li, R. Chu, S. Liao, L. Zhang. Illumination invariant face recognition using near-infrared images. *PAMI*, 2007.

[11] D. Lin and X. Tang, Inter-modality face recognition, *ECCV*, 2006.

[12] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, *CVPR*, 2005

[13] I. V. Meglinski, S. J. Matcher, Quantitative assessment of skin layers absorption and skin reflectance spectra simulation in the visible and near-infrared spectral regions, *Physiological Measurement*, 2002

[14] T. Ojala, M. Pietikäinen and T. Mäenpää. Multiresolution gray scale and rotation invariant texture analysis with local binary patterns, *PAMI*, 2002.

[15] P.J. Phillips, H. Moon, S. A.Rizvi and P. J.Rauss, The FERET evaluation methodology for face-recognition algorithms, *PAMI*, 2000.

[16] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality reduction by locally linear embedding, *Science*, 2000.

[17] X. Tang and X. Wang, Face sketch synthesis and recognition, *ICCV*, 2003.

[18] D. Yi, R. Liu, R. Chu, Z. Lei and S. Z. Li, Face matching from near infrared to visual images, *ICB*, 2008

[19] W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld, Face recognition: a literature survey, *ACM Computing Surveys*, 2003.