Pattern Recognition 44 (2011) 1567–1580

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Kernel discriminant transformation for image set-based face recognition

Wen-Sheng Chu, Ju-Chin Chen, Jenn-Jier James Lien *

Department of Computer Science and Information Engineering, National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan

ARTICLE INFO

Article history:
Received 27 February 2010
Received in revised form
28 January 2011
Accepted 9 February 2011
Available online 16 February 2011

Keywords:
Subspace method
Canonical correlation
Kernel discriminant analysis
Face recognition

ABSTRACT

This study presents a novel kernel discriminant transformation (KDT) algorithm for face recognition based on image sets. As each image set is represented by a kernel subspace, we formulate a KDT matrix that maximizes the similarities of within-kernel subspaces, and simultaneously minimizes those of between-kernel subspaces. Although the KDT matrix cannot be computed explicitly in a high-dimensional feature space, we propose an iterative kernel discriminant transformation algorithm to solve the matrix in an implicit way. Another perspective of similarity measure, namely canonical difference, is also addressed for matching each pair of the kernel subspaces, and employed to simplify the formulation. The proposed face recognition system is demonstrated to outperform existing still-image-based as well as image set-based face recognition methods using the Yale Face database B, Labeled Faces in the Wild and a self-compiled database.

1. Introduction

Automatic face recognition has been an essential requirement in a wide range of computer vision applications, such as human–computer interaction (HCI), content-based image retrieval (CBIR) [1,2], security systems and access control systems [3–5]. However, automatic face recognition is a complex process since the appearance of facial images contains an immense variety of expressions, orientations, lighting conditions, occlusions and so forth. Due to these dramatic variations of facial images, the performance of conventional face recognition systems based on a single testing image [3,6–9] is somewhat limited.

Many researchers have noticed the influence of appearance variations for face recognition, and pursued many studies to tackle this problem. Multiple classifiers [10] and active appearance models (AAM) [11] address facial pose issues. A survey also can be also found in [12]. Illumination compensation [13], quotient image creation [14], and synthesized illuminated exemplars [15] are proposed to tackle the illumination problems. In addition, multiscale facial structure [16] and local contrast enhancement [17] is also explored to recognize faces under varying illumination.

Notwithstanding the contributions of [10,11,13–15,18], a single testing input provides insufficient information to guarantee a reliable recognition performance. Better performance could be obtained from *sets* of testing images since multiple images provide more appearance variances of the input data. Recently, canonical correlation (or principal angles) as a similarity between two image sets has drawn increasing attention. The idea of canonical correlation is to measure the cosine angles between associated basis pairs of linear subspaces that correspond to image sets. Canonical correlation has been an effective representation in capturing image set information [19–23]. However, for nonlinearly distributed patterns, such as facial images with head motions and lighting variations (see Fig. 1 for example), these methods are somewhat limited due to their assumption of linearity.

Since facial images are nonlinearly distributed, this study develops a novel kernel discriminant transformation (KDT) algorithm for image-set based face recognition. The proposed algorithm is based on a canonical difference measure [24]. In spirit, the proposed KDT algorithm bears some resemblances to the DCC method [22] in that they both solve an optimal transformation based on a discriminative criterion and the concept of canonical correlations. However, our method has several significant differences in extending the DCC method to a nonlinear version, i.e., the entire learning process is performed in a high-dimensional feature space. Each input image set is represented as a kernel subspace, instead of a linear subspace, by considering the nonlinearity of KPCA [25].

Since the input here is different from DCC [21], our contribution can be clearly summarized into four-fold: (1) Due to the nonlinearity of facial images, we propose to learn a KDT matrix such that after the transformation, separation within kernel subspaces of the same class are minimized, and at the same time separation between kernel subspaces of different classes are maximized. (2) Although the KDT matrix cannot be explicitly computed in high-dimensional feature space, we develop the KDT algorithm to derive implicit evaluation for the dot products in the

* Corresponding author.
E-mail address: jjlien@csie.ncku.edu.tw (J.J. Lien).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.011

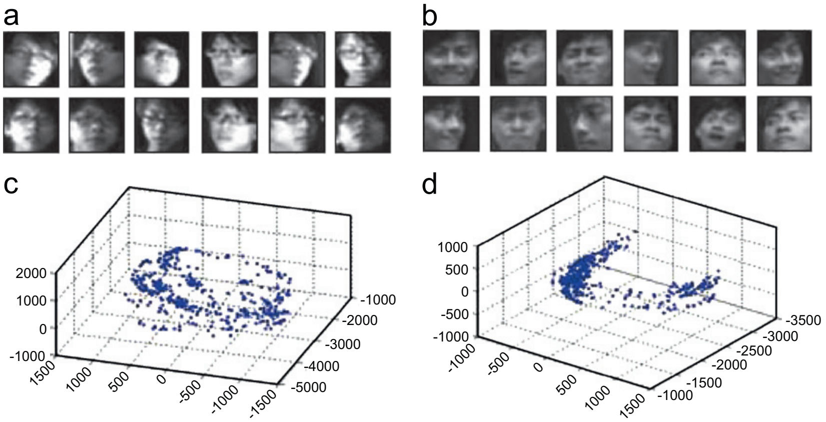

Fig. 1. Typical examples of facial images containing unconstrained head motions with different lighting conditions (a) and facial expressions (b). (c) and (d) show respective nonlinear distribution in the 3D eigenspace.

feature space. (3) We discuss a geometric perspective of canonical correlation, namely *canonical difference*, for measuring the similarity between subspace pairs, and show that the correlation and the difference are in inverse proportion of each other based on simple geometric rules. (4) We provide analysis of computational complexity, bounds, and significance testing for the proposed algorithm.

The rest of this paper is organized as follows: next section reviews previous work related to this paper. An overview of the proposed approach is illustrated in Section 3. Section 4 presents the training procedure, computational analysis and bounds of the proposed KDT algorithm. The testing process is shown in Section 5. We show the experimental results in Section 6, and draw the conclusion in Section 7.

2. Related work

Reviewing the literature, image set-based face recognition approaches mainly fall into two categories: temporal-based and non-temporal approaches. Temporal-based approaches [26–29] recognize human faces by analyzing the connectivity among temporal sequences in video. Non-temporal approaches [20–22,30–36], on the other hand, require no assumption of temporal coherence between facial images, and thus have an advantage that the training database can be arbitrarily expanded instead of being recollected. The proposed method belongs to the later category, therefore we focus on the discussion on non-temporal approaches.

Non-temporal approaches for image set-based face recognition can be further separated into two types [22]: sample-based (nonparametric) and model-based (parametric) methods. In methods of the former type, such as [26], the recognition process involves matching pairwise samples in two image sets, and thus can be time-consuming and sensitive to noise/outliers. By contrast, model-based methods, such as [30,34], assume a preliminary statistical model for each facial image set, and thus requires a strong statistical correlation between the training data and the testing data to ensure satisfactory recognition. To identify that if two facial image sets belong to the same person, the most effective way would be measuring the similarity of the common views of data, which is the idea of canonical correlations.

Yamaguchi et al. propose a set-based face recognition system based on a mutual subspace method (MSM) [19], and yield an improvement compared to image-based approaches that use only one image for testing. Each image set is represented as a linear subspace, and canonical correlations are exploited as the similarity between two image sets. However, MSM does not consider inter-subspace information and therefore limits its discriminative power. Constrained mutual subspace method (CMSM) [20], on the other hand, defines a constrained subspace as differences between all training subspaces, and then measures the canonical correlations in the constrained subspace. Kim et al. [22] report that the performance of CMSM depends on an appropriate dimensionality of the constraint subspace, and propose an alternative scheme to learn a discriminative canonical correlation without specifying dimensionality of target subspace. The recognition performance is shown more robust than both conventional sample-based and model-based methods. Nevertheless, above methods are developed under the linear assumption of input patterns.

Nonlinear extensions of face recognition systems have been widely proposed in [25,37–42]. For these methods, a nonlinear function is applied to map input patterns in the original space to a high-dimensional feature space, where nonlinear input patterns are shown to be more easily classified. The inner products of images in the feature space can be implied by a kernel function in the input space. Regarding the image-based methods, the kernel property is employed in [36] to nonlinearly extend the MSM method [19], denoted as kernel MSM (KMSM). As in the original MSM method, the proposed approach ignores the discriminative information between different classes, and thus restricts its performance. Fukui et al. [31] address the nonlinear version of CMSM in by deriving a representation for the kernel constrained subspace, and justify their performance through an application of 3D object recognition. However, KCMSM inherits the same drawbacks of a appropriate dimensionality for the constrained subspace. In a later study, Fukui and Yamaguchi [32] propose a kernel orthogonal mutual subspace method (KOMSM), where the selection of dimensionality was resolved by orthogonalizing the kernel subspaces before calculating their canonical correlations. Inspired by the above effectiveness of kernel methods, this work proposes a more reliable approach to tackle the problem of image-set based face recognition.

3. Overview of the proposed image set-based face recognition system

Fig. 2 shows the overview of training and testing processes in the proposed image set-based face recognition system. The training process commences by compiling n (typically $n=3$) image sets for each subject. Each image set comprises n_i 20×20-pixel facial images characterized by arbitrary head

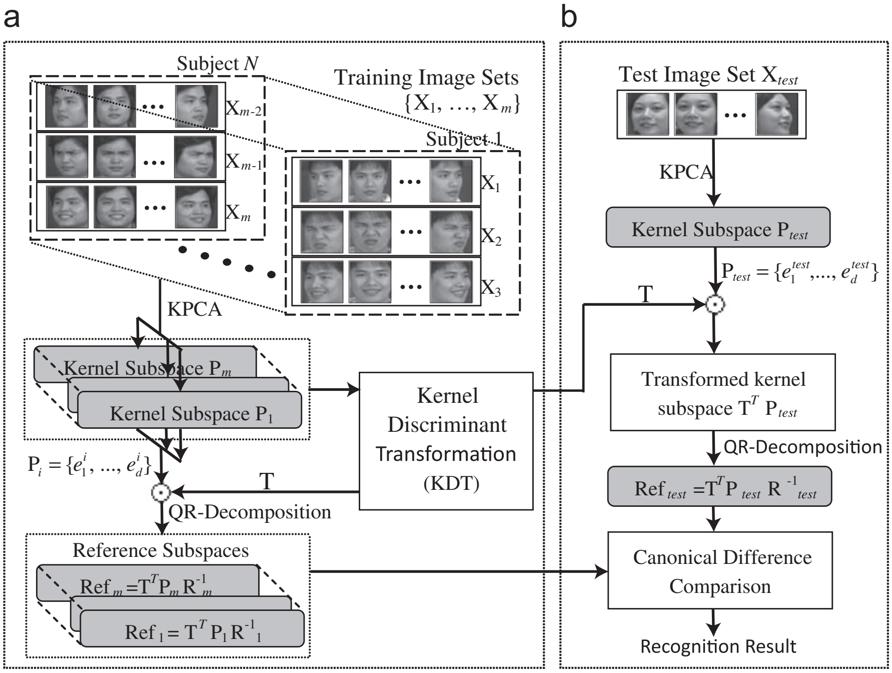
a
b
Subject N
Training Image Sets
$\{X_1, \dots, X_m\}$
X_{m-2}
X_{m-1}
X_m
Subject 1
X_1
X_2
X_3
KPCA
Kernel Subspace P_m
Kernel Subspace P_1
$P_i = \{e_1^i, \dots, e_d^i\}$
T
QR-Decomposition
Kernel Discriminant Transformation (KDT)
Reference Subspaces
$\text{Ref}_m = T^T P_m R^{-1}_m$
$\text{Ref}_1 = T^T P_1 R^{-1}_1$
Test Image Set X_{test}
KPCA
Kernel Subspace P_{test}
$P_{test} = \{e_1^{test}, \dots, e_d^{test}\}$
T
Transformed kernel subspace $T^T P_{test}$
QR-Decomposition
$\text{Ref}_{test} = T^T P_{test} R^{-1}_{test}$
Canonical Difference Comparison
Recognition Result

Fig. 2. Flowchart of the proposed face recognition system: (a) training process, (b) testing process; ⊙ denotes the transformation of the KDT matrix **T** and the kernel subspace $\mathbf{P}_i$.

motions and either lighting condition variations or facial expression variations. The total number of subjects is assumed to be N, and thus the training database contains a total of $m\ (=N\times n)$ image sets.

Since the facial images in the image sets $\mathbf{X}_i\ (i=1,\dots,m)$ produce nonlinear manifolds in the original space (as shown in Fig. 1), we perform KPCA on each $\mathbf{X}_i$ to represent the bases spanning the corresponding kernel subspace $\mathbf{P}_i$ in the high-dimensional feature space $\mathcal{F}$. Then, the proposed kernel discriminant transformation (KDT) algorithm is employed to obtain an optimal transformation matrix T such that the transformed kernel subspace $\mathbf{T}^T\mathbf{P}_i$ $(i=1,\dots,m)$ give maximal correlation between kernel subspaces related the same subject and minimal correlation between those related to different subjects. Note that what we explicitly compute in the KDT algorithm is the transformed kernel subspaces $\mathbf{T}^T\mathbf{P}_i$ instead of the KDT matrix **T**. A more detailed derivation is shown in the following sections that **T** can be used for evaluating the inner products of $\mathbf{T}^T\mathbf{P}_i$ in $\mathcal{F}$. We then provide another perspective of the similarity measure of two image sets by considering the difference between associated canonical vectors [22] of two subspace. To satisfy the orthonormal condition of canonical vectors, a QR-decomposition operation is performed on each transformed subspace $\mathbf{T}^T\mathbf{P}_i$ to produce a corresponding reference subspace Ref_i for each image set $\mathbf{X}_i$.

In the testing process, as shown in Fig. 2(b), a reference subspace Ref_{test} corresponding to the input image $\mathbf{X}_{test}$ is obtained using the same procedure as in the training process. The similarity between reference subspaces Ref_{test} and Ref_i in the training database is evaluated in terms of the canonical differences between them. The training reference subspace Ref_i yielding the smallest canonical difference is then returned as the recognition result.

4. Training process

In this section, we discuss the details of the proposed KDT algorithm and the similarity measure of canonical difference in the training process. The training process commences by utilizing KPCA to represent a kernel subspace $\mathbf{P}_i$ for each image set $\mathbf{X}_i$. Then, we define a transformation matrix **T** in $\mathcal{F}$, denoted as *the KDT matrix* hereafter for convenience, based on Fisher's linear discriminant and the canonical difference as a similarity measure between each pair of kernel subspaces. The idea of **T** is to transform all kernel subspaces such that the transformed kernel subspaces $\mathbf{T}^T\mathbf{P}_i$ $(i=1,\dots,m)$ give maximal correlation between kernel subspaces related to the same subject and minimal correlation between those related to different subjects. As **T** cannot be computed explicitly in $\mathcal{F}$, we reformulate Fisher's linear discriminant as kernel Fisher's discriminant (KFD) to solve the matrix α, which represents the coefficients of the bases of **T**. The entire procedure of finding the optimal coefficient matrix α is referred as the *kernel discriminant transformation* (*KDT*) *algorithm*. By examining the proposed KDT algorithm, we offer the analysis of the bounds of the number of bases for **T** and the complexity of the KDT algorithm. Finally, a reference subspace of each kernel subspace is calculated for computational convenience.

In the following discussion, the set of training data $\{\mathbf{X}_1,\dots,\mathbf{X}_m\}$ is assumed to contain N subjects with n image sets per subject, i.e. a total of $m\ (=N\times n)$ image sets in the training database. Each image set $\mathbf{X}_i = [\mathbf{x}_1,\dots,\mathbf{x}_{n_i}]$ contains n_i 20 × 20-pixel facial images.

4.1. Kernel subspace creation using KPCA

As shown in Fig. 1, the facial images with different lighting conditions or facial expressions are nonlinearly distributed in the original space. Hence, their distribution cannot be well represented using a linear subspace method such as PCA. In this study, we employ the kernel PCA (KPCA) method [25] to represent a kernel subspace $\mathbf{P}_i$ for each training image set $\mathbf{X}_i$. It was addressed in [42] that the use of kernel subspaces is capable of extracting an abundance of nonlinear features from facial or object images. KPCA, therefore, provides a nonlinear solution for automatic face recognition systems to identify facial images characterized by wide variations in poses, facial expressions or illumination.

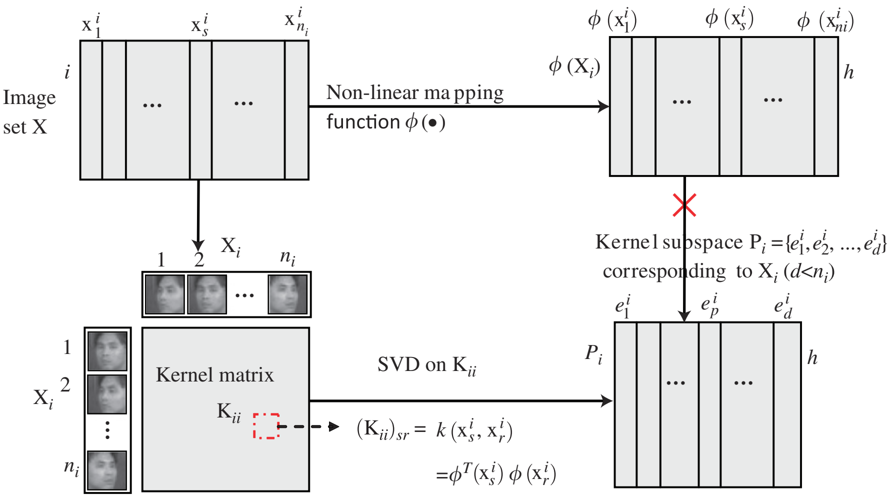
x_1^i
x_s^i
$x_{n_i}^i$
i
Image set X
Non-linear mapping function $\phi(\bullet)$
$\phi(x_1^i)$
$\phi(x_s^i)$
$\phi(x_{ni}^i)$
$\phi(X_i)$
h
Kernel subspace $P_i=\{e_1^i, e_2^i, \ldots, e_d^i\}$
corresponding to X_i ($d<n_i$)
e_1^i
e_p^i
e_d^i
X_i
1
2
n_i
P_i
h
Kernel matrix
K_{ii}
SVD on K_{ii}
$(K_{ii})_{sr}= k(x_s^i, x_r^i)$
$=\phi^T(x_s^i)\,\phi(x_r^i)$
X_i
1
2
n_i

Fig. 3. Kernel subspace representation using KPCA. Rather than directly applying a nonlinear mapping function $\phi(\cdot)$ to each facial image $\mathbf{x}_s^i$ of the image set $\mathbf{X}_i$, the kernel function $k(\mathbf{x}_s^i,\mathbf{x}_r^i)$ is used to generate the kernel matrix $\mathbf{K}_{ii}$ for $\mathbf{X}_i$. The corresponding kernel subspace $\mathbf{P}_i$ is then obtained as the span of the eigenvectors $\{e_p^i\}_{p=1}^d$.

As shown in Fig. 3, each image set $\mathbf{X}_i$ ($i=1,\ldots,m$) in the original input space is mapped into a high-dimensional feature space $\mathcal{F}$ using the following nonlinear mapping function:

$$\phi : \{\mathbf{X}_i, \ldots, \mathbf{X}_m\} \to \{\phi(\mathbf{X}_i), \ldots, \phi(\mathbf{X}_m)\}, \tag{1}$$

where m is the number of training image sets. In practice, the dimensionality of $\mathcal{F}$, denoted as h, can be huge or possibly infinite, and thus performing calculations in $\mathcal{F}$ is computationally complex and expensive. Using the "kernel trick", the dot products $\phi^T(\mathbf{x}_r)\phi(\mathbf{x}_s)$ in $\mathcal{F}$ can be expressed in terms of a kernel function $k(\mathbf{x}_r,\mathbf{x}_s)$ such that calculation in $\mathcal{F}$ can be easily performed in terms of dot products instead of direct use of the mapped images $\phi(\mathbf{x})$. The training process proposed in this study utilizes the Gaussian radial basis function (RBF) kernel [25,31,38,41,42]:

$$k(\mathbf{x}_r,\mathbf{x}_s) = \exp\left(\frac{-\|\mathbf{x}_r-\mathbf{x}_s\|^2}{\sigma^2}\right), \tag{2}$$

where $\mathbf{x}_r$ and $\mathbf{x}_s$ are image vectors of gray-value and σ is the variance of the pixel intensity. We then compute an $n_i \times n_j$ kernel matrix $\mathbf{K}_{ij}$ according to the kernel function $k(\mathbf{x}_r,\mathbf{x}_s)$ [25], which gives each element of $\mathbf{K}_{ij}$ as follows:

$$(\mathbf{K}_{ij})_{rs} = \phi^T(\mathbf{x}_r^i)\phi(\mathbf{x}_s^j) = k(\mathbf{x}_r^i,\mathbf{x}_s^j), \tag{3}$$

where $\mathbf{x}_r^i$ is the r-th image in the image set $\mathbf{X}_i$ and $\mathbf{x}_s^j$ is the s-th image in X_j; $r=1,\ldots,n_i$, $s=1,\ldots,n_j$.

As shown in Fig. 3, the kernel subspace $\mathbf{P}_i$ for each image set $\mathbf{X}_i$ in $\mathcal{F}$ is obtained by performing KPCA using the kernel matrix $\mathbf{K}_{ii}$, i.e. $j=i$. Note that $\phi(\mathbf{X}_i)$ in this KPCA case is not centered in $\mathcal{F}$ [31,33,36] as the linear subspaces obtained from the correlation matrix in these studies [19–21]. By applying singular value decomposition (SVD) to $\mathbf{K}_{ii}$, we can obtain

$$\mathbf{K}_{ii} = \mathbf{a}_i \Gamma \mathbf{a}_i^T, \tag{4}$$

where $\mathbf{a}_i$ denotes an $n_i \times n_i$ matrix of eigenvectors and Γ denotes an $n_i \times n_i$ diagonal matrix of eigenvalues. In accordance with the theory of reproducing kernels, e_p^i (the p-th eigenvector of $\mathbf{P}_i$) can be expressed as the linear combination of mapped images:

$$e_p^i = \sum_{s=1}^{n_i} \mathbf{a}_{sp}^i \phi(\mathbf{x}_s^i). \tag{5}$$

The coefficient $\mathbf{a}_{sp}^i$ is the s-th component of the eigenvector corresponding to the p-th largest eigenvalue of $\mathbf{K}_{ii}$. Denoting the number of bases of $\mathbf{P}_i$ as d, $\mathbf{P}_i$ can be represented as the span of the eigenvectors $\{e_p^i\}_{p=1}^d$, i.e. $\mathbf{P}_i = [e_1^i, \ldots, e_d^i]$. In the proposed training process, a kernel subspace $\mathbf{P}_i \in \mathbb{R}^{h\times d}$ is exploited to represent the nonlinear manifold for each image set $\mathbf{X}_i$.

4.2. The kernel discriminant transformation (KDT) algorithm

With kernel subspaces $\mathbf{P}_i$ standing for image sets $\mathbf{X}_i$ ($i=1,\ldots,m$), a kernel discriminant transformation (KDT) algorithm is then proposed based on the kernel Fisher discriminant formulation [40] to obtain an optimal representation for the KDT matrix $\mathbf{T}$. The idea of the matrix $\mathbf{T}$ is to transform each kernel subspace such that the correlation between kernel subspaces of different subjects is minimized, while that between kernel subspaces of the same subject is maximized. In order to measure the correlation between two subspaces, we present a different perspective from canonical correlation [22], called *canonical difference* in this paper. The formulation for the KDT matrix $\mathbf{T}$ is presented in Section 4.2.1. While the explicit solution of $\mathbf{T}$ is intractable, we provide an optimized solution through the KDT algorithm in Section 4.2.2. Finally, the bounds of the number of bases for $\mathbf{T}$ and the complexity of the proposed algorithm is analyzed in Section 4.2.3 and 4.2.4, respectively.

4.2.1. Formulation

The main idea of the KDT algorithm is to find an $h \times w$ KDT matrix $\mathbf{T}$ which maximizes the correlation of within-subspaces and minimizes that of between-subspaces, such that the relationship between the entire kernel subspaces in the training database can be established. The KDT matrix $\mathbf{T}$ consists w bases in the high-dimensional feature space $\mathcal{F}$, where the dimensionality is h. Let $\mathbf{T}^T\mathbf{P}_i$ denote the kernel subspace $\mathbf{P}_i$ transformed by $\mathbf{T}$. To satisfy the definition of canonical correlation [22,31,33,36], each $\mathbf{T}^T\mathbf{P}_i$ is required to be unitary orthogonal bases. In this study, the normalization process is performed by applying a QR-decomposition operation to each instance of $\mathbf{T}^T\mathbf{P}_i$:

$$\mathbf{T}^T\mathbf{P}_i = \mathbf{Q}_i\mathbf{R}_i, \tag{6}$$

where $\mathbf{Q}_i$ is a $w \times d$ orthonormal matrix and $\mathbf{R}_i$ is a $d \times d$ invertible upper triangular matrix. Rewriting $\mathbf{Q}_i$ corresponding to $\mathbf{T}^T\mathbf{P}_i$ as

$$\mathbf{Q}_i = \mathbf{T}^T\mathbf{P}_i\mathbf{R}_i^{-1} \tag{7}$$

the similarity measure of canonical correlation between $\mathbf{Q}_i$ and $\mathbf{Q}_j$ can then be expressed in the form

$$\Lambda = \Phi_{ij}^T\mathbf{Q}_i^T\mathbf{Q}_j\Phi_{ji} = \mathbf{C}_i^T\mathbf{C}_j \quad \text{s.t. } \Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_n), \tag{8}$$

where Φ_{ij} and Φ_{ji} are eigenvectors and $\{\sigma_1,\ldots,\sigma_n\}$ are eigenvalues of the SVD on $\mathbf{Q}_i^T\mathbf{Q}_j$. The canonical subspaces are defined as $\mathbf{C}_i=\mathbf{Q}_i\Phi_{ij}=[\mathbf{u}_1,\ldots,\mathbf{u}_d]$ and $\mathbf{C}_j=\mathbf{Q}_j\Phi_{ji}=[\mathbf{v}_1,\ldots,\mathbf{v}_d]$, respectively, where Φ_{ij} and Φ_{ji} are rotation matrices [22] of $\mathbf{Q}_i$,$\mathbf{Q}_j$, and $\mathbf{u}_k$,$\mathbf{v}_k$ (k=1,…,d) are the canonical vectors in $\mathcal{F}$. For illustration purpose, we show in Fig. 4 the eigenvectors, canonical vectors, and corresponding differences in image space, while in our work, the image space is mapped to a high dimensional feature space $\mathcal{F}$. As shown in Fig. 4, canonical vectors can more effectively capture the facial expression and lighting condition variations in each image set than the eigenvectors.

From a geometric perspective, the difference between two unitary vectors is proportional to the angle between them, as shown in Fig. 5. Since canonical vectors $\mathbf{u}_k$, $\mathbf{v}_k$ in $\mathbf{C}_i$, $\mathbf{C}_j$ are unitary vectors, the difference $\mathbf{d}_k$ between $\mathbf{u}_k$ and $\mathbf{v}_k$, i.e. $\mathbf{d}_k=\mathbf{u}_k-\mathbf{v}_k$, is proportional to their included angle. Similar to canonical correlation, the canonical difference is used to measure the similarity between two unitary orthogonal subspaces from a different viewpoint. Specifically, the greater the similarity between the two subspaces, the smaller the value of the canonical difference. The canonical difference is defined in terms of the squared sum of distances $\|\mathbf{d}_k\|^2$ between all the vectors in canonical subspaces $\mathbf{C}_i$ and $\mathbf{C}_j$:

$$\mathit{CanonicalDiff}(i,j)=\sum_{k=1}^{d}\|\mathbf{u}_k-\mathbf{v}_k\|^2=\mathrm{trace}((\mathbf{C}_i-\mathbf{C}_j)^T(\mathbf{C}_i-\mathbf{C}_j)). \quad (9)$$

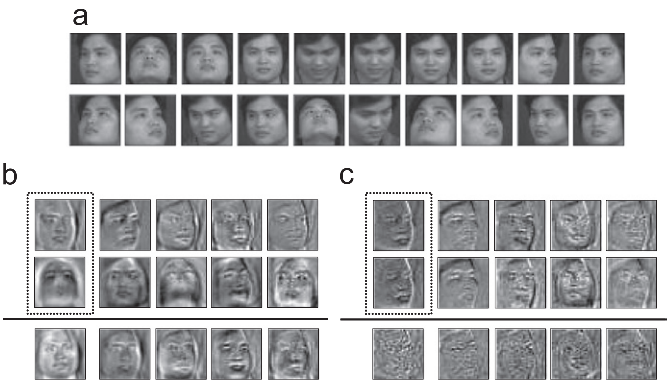

Fig. 4. Conceptual illustration in image space. Facial images in each row represent (a) an image set, (b) eigenvectors and corresponding differences, and (c) canonical vectors and corresponding differences. In (b) and (c), the first two rows show the first five eigenvectors (or principal components) and first five canonical vectors, while the third row shows the difference between them.

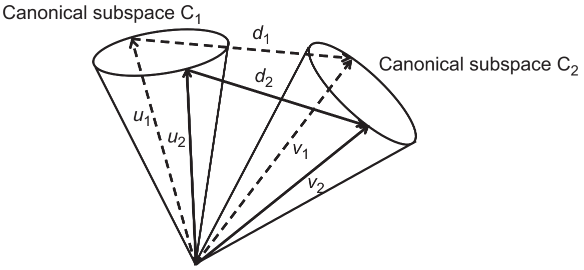

Fig. 5. Conceptual illustration of canonical differences, in which the canonical subspaces $\mathbf{C}_1$, $\mathbf{C}_2$ of image sets $\mathbf{X}_1$, $\mathbf{X}_2$ are spanned by the orthonormal vectors $[\mathbf{u}_1,\ldots,\mathbf{u}_d]$ and $[\mathbf{v}_1,\ldots,\mathbf{v}_d]$, respectively. According to geometrical principles, the difference $\mathbf{d}_i$ between $\mathbf{u}_i$ and $\mathbf{v}_i$, i.e. $\mathbf{d}_i=\mathbf{v}_i-\mathbf{u}_i$, is proportional to the angle between them. In the proposed face recognition system, the canonical difference is used to measure the similarity between kernel subspaces. Specifically, the greater the similarity between $\mathbf{X}_1$ and $\mathbf{X}_2$, the smaller the value of the canonical difference is.

It is observed from Eq. (9) that when two canonical subspaces are separated by a small distance, the trace term yields a small value for the sum of the diagonal elements. With the definition of canonical differences, we now begin to develop the KDT formulation: we apply the definition of the canonical subspaces given above and rewrite Eq. (9) as

$$\mathit{CanonicalDiff}(i,j)=\mathrm{tr}((\mathbf{Q}_i\Phi_{ij}-\mathbf{Q}_j\Phi_{ji})^T(\mathbf{Q}_i\Phi_{ij}-\mathbf{Q}_j\Phi_{ji})). \quad (10)$$

Substituting Eq. (7) into Eq. (10), the canonical difference can be expressed as

$$\mathit{CanonicalDiff}(i,j)=\mathrm{tr}(\mathbf{T}^T(\mathbf{P}_i\Phi'_{ij}-\mathbf{P}_j\Phi'_{ji})(\mathbf{P}_i\Phi'_{ij}-\mathbf{P}_j\Phi'_{ji})^T\mathbf{T}), \quad (11)$$

where $\Phi'_{ij}=\mathbf{R}_i^{-1}\Phi_{ij}$ and $\Phi'_{ji}=\mathbf{R}_j^{-1}\Phi_{ji}$. The KDT matrix $\mathbf{T}$ can then be formulated as the problem of maximizing the ratio of the canonical differences of the between-subspaces to those of the within-subspaces. In other words, finding the KDT matrix $\mathbf{T}$ can be regarded as the optimization of Fisher's linear discriminant in $\mathcal{F}$, i.e.

$$\mathbf{T}=\underset{\mathbf{T}}{\mathrm{argmax}}\frac{\sum_{i=1}^{m}\sum_{l\in B_i}\mathit{CanonicalDiff}(i,l)}{\sum_{i=1}^{m}\sum_{k\in W_i}\mathit{CanonicalDiff}(i,k)}$$

$$=\underset{\mathbf{T}}{\mathrm{argmax}}\frac{\mathrm{trace}(\mathbf{T}^T\mathbf{S}_b\mathbf{T})}{\mathrm{trace}(\mathbf{T}^T\mathbf{S}_w\mathbf{T})}, \quad (12)$$

where the between-scatter matrix $\mathbf{S}_b$ is given by

$$\mathbf{S}_b=\sum_{i=1}^{m}\sum_{l\in B_i}(\mathbf{P}_i\Phi'_{il}-\mathbf{P}_l\Phi'_{li})(\mathbf{P}_i\Phi'_{il}-\mathbf{P}_l\Phi'_{li})^T \quad (13)$$

and the within-scatter matrix $\mathbf{S}_w$ by

$$\mathbf{S}_w=\sum_{i=1}^{m}\sum_{k\in W_i}(\mathbf{P}_i\Phi'_{ik}-\mathbf{P}_l\Phi'_{ki})(\mathbf{P}_i\Phi'_{ik}-\mathbf{P}_l\Phi'_{ki})^T. \quad (14)$$

Note that the class label of the image set $\mathbf{X}_i$ is denoted by c_i, while the sets $B_i=\{l|c_l\neq c_i\}$, $W_i=\{k|c_k=c_i\}$ denote the class labels of the between and within subspaces, respectively. The iterative procedure performed to optimize Eq. (12) is summarized in Fig. 6.

4.2.2. Optimization

Since we have formulated the KDT problem as a Fisher's linear discriminant in high-dimensional feature space $\mathcal{F}$, the solution of the KDT matrix $\mathbf{T}$ is intractable because the formulation in Eq. (12) depends on the variables $\mathbf{P}$, Φ'_{ij} and most importantly on $\mathbf{T}$ as well. This means there is no explicit expression for $\mathbf{T}$ in terms of the other two variables. Hence, an iteration procedure is needed to obtain the optimal KDT matrix $\mathbf{T}$. The procedure commences by obtaining the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$, which are then taken as the input to the following loops. Note that in the scatter matrix formulations of Eqs. (13) and (14), the kernel functions cannot be substituted into the dot products $\phi^T(\mathbf{x}_r)\phi(\mathbf{x}_s)$. The KDT matrix $\mathbf{T}$, therefore, cannot be solved directly by the eigen-decomposition problem of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

Here, we offer an alternative solution by reformulating Fisher's linear discriminant in $\mathcal{F}$ given in Eq. (12) to the form of kernel Fisher discriminant (KFD). Assuming the number of training images is M, i.e. $M=\sum_{i=1}^{m}n_i$, we apply the theory of reproducing kernels to represent the vectors $\{\mathbf{t}_q\}_{q=1}^{w}\in\mathbf{T}^{h\times w}$ as the span of all the mapped training images:

$$\mathbf{t}_q=\sum_{u=1}^{M}\alpha_{uq}\phi(\mathbf{x}_u), \quad (15)$$

where α_{uq} is an element of the $M\times d$ coefficient matrix α. Rather than directly solving Eq. (12), we instead compute the

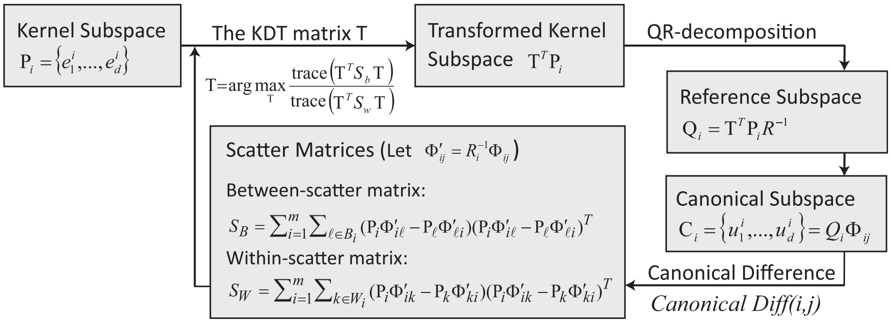
Kernel Subspace
$P_i=\{e_1^i,\ldots,e_d^i\}$
The KDT matrix T
$T=\arg\max_T \frac{\mathrm{trace}(T^TS_bT)}{\mathrm{trace}(T^TS_wT)}$
Transformed Kernel Subspace T^TP_i
QR-decomposition
Reference Subspace
$Q_i=T^TP_iR^{-1}$
Canonical Subspace
$C_i=\{u_1^i,\ldots,u_d^i\}=Q_i\Phi_{ij}$
Scatter Matrices (Let $\Phi'_{ij}=R_i^{-1}\Phi_{ij}$)
Between-scatter matrix:
$S_B=\sum_{i=1}^m\sum_{\ell\in B_i}(P_i\Phi'_{i\ell}-P_\ell\Phi'_{\ell i})(P_i\Phi'_{i\ell}-P_\ell\Phi'_{\ell i})^T$
Within-scatter matrix:
$S_W=\sum_{i=1}^m\sum_{k\in W_i}(P_i\Phi'_{ik}-P_k\Phi'_{ki})(P_i\Phi'_{ik}-P_k\Phi'_{ki})^T$
Canonical Difference
Canonical Diff(i,j)

Fig. 6. Iterative optimization of the KDT matrix **T**. As shown, the solution procedure commences by applying an initialized **T** to each kernel subspace $\mathbf{P}_i$. The reference subspace $\mathbf{Q}_i$, canonical subspace $\mathbf{C}_i$ and Φ'_{ij} are then computed following a QR-decomposition operation performed on $\mathbf{T}^T\mathbf{P}_i$. The between-scatter matrix $\mathbf{S}_B$ and within-scatter matrix $\mathbf{S}_W$ are then obtained in accordance with the canonical difference similarity measure. Taking the updated values of $\mathbf{S}_B$ and $\mathbf{S}_W$ as the new inputs, the procedure is repeated iteratively to establish the optimized KDT matrix **T**.

transformed kernel subspace $\mathbf{T}^T\mathbf{P}_i$ as

$$(\mathbf{T}^T\mathbf{P}_i)_{qp}=\sum_{u=1}^{M}\sum_{s=1}^{n_i}\alpha_{uq}\mathbf{a}_{sp}^i k(\mathbf{x}_u,\mathbf{x}_s^i), \tag{16}$$

where $\mathbf{a}_{sp}^i$ is obtained from Eq. (4) and $k(\mathbf{x}_u,\mathbf{x}_s^i)$ is computed using the u-th training image $\mathbf{x}_u$ and the s-th image of the image set $\mathbf{X}_i$. Given $\mathbf{T}^T\mathbf{P}_i$, the orthonormal bases matrix $\mathbf{Q}_i$ and $\mathbf{Q}_j$ are obtained by the QR-decomposition operation (as shown in Eq. (6)) on $\mathbf{T}^T\mathbf{P}_i$. The rotation matrices Φ_{ij} and Φ_{ji} are computed through SVD on $\mathbf{Q}_i^T\mathbf{Q}_j$. Then, Φ'_{ij} and Φ'_{ji} in Eq. (11) are obtained from $\mathbf{R}_i^{-1}\Phi_{ij}$ and $\mathbf{R}_j^{-1}\Phi_{ji}$, respectively. By substituting e_p^i in Eq. (5) into the bases of $\mathbf{P}_i$, the rotated kernel subspace $\tilde{\mathbf{P}}_{ij}$ for $\mathbf{P}_i$ can be expressed as

$$\tilde{\mathbf{P}}_{ij}=\mathbf{P}\Phi'_{ij}=[\tilde{e}_1^{ij},\ldots,\tilde{e}_d^{ij}], \tag{17}$$

where the p-th basis vector is computed by

$$\tilde{e}_p^{ij}=\sum_{r=1}^{d}\sum_{s=1}^{n_i}\mathbf{a}_{sr}^i\Phi_{ij}^{\prime rp}\phi(\mathbf{x}_s^i). \tag{18}$$

Through the multiplication of Eq. (15) and Eq. (18), it can be shown that $\mathbf{T}^T\tilde{\mathbf{P}}_{ij}=\alpha\mathbf{Z}_{ij}$, where each element of $\mathbf{Z}_{ij}$ has the form

$$(\mathbf{Z}_{ij})_{up}=\sum_{r=1}^{d}\sum_{s=1}^{n_i}\mathbf{a}_{sr}^i\Phi_{ij}^{\prime rp}k(\mathbf{x}_u,\mathbf{x}_s^i), \tag{19}$$

where u=1,…,M and p=1,…,d. Note that $\mathbf{Z}_{ij}$ is tractable as a result of replacing the dot products $\phi^T(\mathbf{x}_r)\phi(\mathbf{x}_s)$ with $k(\mathbf{x}_i,\mathbf{x}_j)$ in Eq. (19). Applying the definitions of $\mathbf{t}_q$ in Eq. (15), $\tilde{e}_p^{ij}$ in Eq. (18) and $\mathbf{Z}_{ij}$ in Eq. (19), the denominator of Eq. (12) can be derived as follows:

$$\begin{aligned}\mathbf{T}^T\mathbf{S}_w\mathbf{T}&=\sum_{i=1}^{m}\sum_{k\in W_i}\sum_{r=1}^{d}[\mathbf{T}^T(\tilde{e}_r^{ki}-\tilde{e}_r^{ik})(\tilde{e}_r^{ki}-\tilde{e}_r^{ik})^T\mathbf{T}]\\&=\sum_{i=1}^{m}\sum_{k\in W_i}\sum_{r=1}^{d}[\mathbf{T}^T\tilde{e}_r^{ki}\tilde{e}_r^{ki^T}\mathbf{T}-\mathbf{T}^T\tilde{e}_r^{ki}\tilde{e}_r^{ik^T}\mathbf{T}-\mathbf{T}^T\tilde{e}_r^{ik}\tilde{e}_r^{ki^T}\mathbf{T}+\mathbf{T}^T\tilde{e}_r^{ik}\tilde{e}_r^{ik^T}\mathbf{T}]\\&=\sum_{i=1}^{m}\sum_{k\in W_i}[\alpha^T\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\alpha-\alpha^T\mathbf{Z}_{ki}\mathbf{Z}_{ik}^T\alpha-\alpha^T\mathbf{Z}_{ik}\mathbf{Z}_{ki}^T\alpha+\alpha^T\mathbf{Z}_{ik}\mathbf{Z}_{ik}^T\alpha]\\&=\sum_{i=1}^{m}\sum_{k\in W_i}[\alpha^T(\mathbf{Z}_{ki}-\mathbf{Z}_{ik})(\mathbf{Z}_{ki}-\mathbf{Z}_{ik})^T\alpha]=\alpha^T\mathbf{U}\alpha,\end{aligned} \tag{20}$$

where α is an $M\times d$ coefficient matrix, and $U=\sum_{i=1}^m\sum_{k\in W_i}(\mathbf{Z}_{ki}-\mathbf{Z}_{ik})(\mathbf{Z}_{ki}-\mathbf{Z}_{ik})^T$ is an $M\times M$ within-scatter matrix. Utilizing a similar procedure to that used to derive Eq. (20), the numerator of Eq. (12) can be rewritten as

$$\mathbf{T}^T\mathbf{S}_b\mathbf{T}=\alpha^T\mathbf{V}\alpha, \tag{21}$$

where $\mathbf{V}=\sum_{i=1}^m\sum_{l\in B_i}(\mathbf{Z}_{li}-\mathbf{Z}_{il})(\mathbf{Z}_{li}-\mathbf{Z}_{il})^T$ is an $M\times M$ between-scatter matrix. Combining Eqs. (20) with Eq. (21), the problem of optimizing Fisher's linear discriminant given in Eq. (12) can be reformulated as the problem of maximizing the following KFD formulation:

$$J(\alpha)=\frac{\mathrm{trace}(\alpha^T\mathbf{V}\alpha)}{\mathrm{trace}(\alpha^T\mathbf{U}\alpha)}. \tag{22}$$

As in the problem encountered by the multi-dimensional Fisher's linear discriminant, α can be found by solving the eigenvectors corresponding to the descending eigenvalues of $\mathbf{U}^{-1}\mathbf{V}$. Although several effective methods [43] have been proposed for solving this problem in the case that $\mathbf{U}$ is not invertible, here we keep it simple by adding a small value μ ($\mu=0.001$) to its diagonal terms [39,40], i.e.

$$\mathbf{U}_\mu=\mathbf{U}+\mu\mathbf{I}. \tag{23}$$

Thus, the inverse $\mathbf{U}_\mu^{-1}$ is guaranteed to exist due to the resulting non-singularity of matrix $\mathbf{U}_\mu$. Finally, the solution of α is obtained to represent the coefficients of the bases of **T** by explicitly computing the leading eigenvectors of $\mathbf{U}_\mu^{-1}\mathbf{V}$. The iterative procedure of the proposed KDT algorithm is summarized in Fig. 7.

The KDT problem is based on the resolution of maximizing the kernel Fisher's discriminant (KFD) formulation: $J(\alpha)=\mathrm{trace}(\alpha^T\mathbf{V}\alpha)/\mathrm{trace}(\alpha^T\mathbf{U}\alpha)$. Our possible approach for solving this problem is to use a 2-step alternative optimization algorithm, followed for example in a different context [44]. The first step of this solution would consist in solving the rotation matrix Φ_{ij} while considering that the $M\times w$ matrix α is fixed. Then, the second step consists in updating α through solving the KFD formulation toward the maximum of the objective function $J(\alpha)$. Once α has been obtained by calculating the eigenvectors of $\mathbf{U}^{-1}\mathbf{V}$ corresponding to w greatest eigenvalues, the rotation matrix Φ_{ij} for the next iteration can be computed. This algorithm stops when an iteration yields a change less than ε in the objective value. Although we do not provide a mathematical proof of convergence of the proposed optimization algorithm, the experiments and analysis have been discussed in Section 6.2.1 to confirm this property.

4.2.3. Bounds of the number of bases for the KDT matrix

The KDT algorithm solves an optimal coefficient matrix for representing the KDT matrix **T**. Given the KFD formulation in Eq. (22), the solution to α is restricted by the rank of matrices **U** and **V**. According to the definition of matrix **Z** in Eq. (19) where rank(**Z**) = d, the rank of matrices **U** and **V** are bounded by

$$d\le\mathrm{rank}(\mathbf{U})\le\min(M,m\times(n-1)\times d/2) \tag{24}$$

and

$$d\le\mathrm{rank}(\mathbf{V})\le\min(M,m\times(m-n)\times d/2), \tag{25}$$

where n is the number of training image sets for each subject, N is the subject (class) number and $m=N\times n$ is the total number of training image sets. To solve $J(\alpha)$ by eigenvectors corresponding to

```
Input: Training image sets $\{\mathbf{X}_i, \ldots, \mathbf{X}_m\}$
Output: A $M \times w$ matrix $\alpha$, which can be applied to $\mathbf{t}_q = \sum_{u=1}^{M} \alpha_{uq}\phi(\mathbf{x}_u)$, $q = 1, \ldots, w$ to
        obtain $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_w]$

1 Given an initial value of $\alpha$: $\alpha^0 \leftarrow \text{rand}_{M\times w}$ and compute $J(\alpha^0)$
  foreach $\mathbf{X}_i (i = 1, \ldots, m)$ do SVD on $\mathbf{K}_{ii}$ to obtain $\mathbf{a}_i$: $\mathbf{K}_{ii} = \mathbf{a}_i \Gamma \mathbf{a}_i^T$
  repeat
2     foreach image set $\mathbf{X}_i$ do
3         Obtain $\mathbf{T}^T\mathbf{P}_i$ by: $\left(\mathbf{T}^T\mathbf{P}_i\right)_{qp} = \sum_{u=1}^{M} \sum_{s=1}^{n_i} \alpha_{uq}\mathbf{a}_{sp}^i k(\mathbf{x}_u, \mathbf{x}_s^i)$
          Obtain $\mathbf{Q}_i$ and $\mathbf{R}_i$ by QR-decomposition of $\mathbf{T}^T\mathbf{P}_i$: $\mathbf{T}^T\mathbf{P}_i = \mathbf{Q}_i\mathbf{R}_i$
4     end
5     foreach pairs of image sets $\mathbf{X}_i$ and $\mathbf{X}_j$ do
6         Obtain $\Phi_{ij}$ by SVD of $\mathbf{Q}_i^T\mathbf{Q}_j$: $\mathbf{Q}_i^T\mathbf{Q}_j = \Phi_{ij}\Lambda\Phi_{ji}^T$ Compute $\Phi'_{ij}$: $\Phi'_{ij} = \mathbf{R}_i^{-1}\Phi_{ij}$
          Compute $\mathbf{Z}_{ij}$: $\left(\mathbf{Z}_{ij}\right)_{up} = \sum_{r=1}^{d} \sum_{s=1}^{n_i} \mathbf{a}_{sr}^i \Phi_{ij}^{\prime rp} k(\mathbf{x}_u, \mathbf{x}_s^i)$
7     end
8     Compute $\mathbf{U}$: $\mathbf{U} = \sum_{i=1}^{m} \sum_{k \in W_i} (\mathbf{Z}_{ki} - \mathbf{Z}_{ik})(\mathbf{Z}_{ki} - \mathbf{Z}_{ik})^T$ Compute $\mathbf{V}$: $\mathbf{V} =$
      $\sum_{i=1}^{m} \sum_{l \in B_i} (\mathbf{Z}_{li} - \mathbf{Z}_{il})(\mathbf{Z}_{li} - \mathbf{Z}_{il})^T$ Obtain updated $\{\alpha_p\}_{p=1}^{w}$ as the first $w$ eigenvectors
      of $\mathbf{U}^{-1}\mathbf{V}$: $\alpha \leftarrow [\alpha_1, \ldots, \alpha_w]$
9 until $\left|J(\alpha^\eta) - J(\alpha^{\eta-1})\right| < \varepsilon$;
```

Fig. 7. The proposed kernel discriminant transformation (KDT) algorithm (n_i: the number of images per image set; d: the number of bases of kernel subspace $\mathbf{P}_i$; w: the number of bases of the KDT matrix $\mathbf{T}$; m: the number of training image sets; M: the number of training images.)

the largest eigenvalues of the $M \times M$ matrix $\mathbf{U}^{-1}\mathbf{V}$ [40], we must ensure that $\mathbf{U}$ is invertible. According to Eq. (24), if matrix $\mathbf{U}$ is full rank of M, i.e. $m \times (n-1) \times d/2 \geq M$, then the inverse of $\mathbf{U}$ exists. Assume that each image set consists of the same number of images n_i, this inequality can be rewritten by $m = N \times n$ and obtain

$$Nd \times n(n-2)/2 \geq M = N \times n \times n_i. \tag{26}$$

Eliminating the common terms on both sides of the equal mark, it can be simplified as $d(n-1) \geq 2n_i$. In this study, the number of bases in each kernel subspace, d, is set to the same value as n_i, which gives

$$n \geq 3. \tag{27}$$

Thus, the existence of $\mathbf{U}^{-1}$ can be guaranteed by the inequality that the number of training image sets per subject is greater than or equal to three. On the other hand, if matrix $\mathbf{U}$ is not full rank of M, a small value μ can be added to the diagonal terms of $\mathbf{U}$ [39,40] as Eq. (23).

By replacing m with $N \times n$ and $n=3$, we can observe that the number of bases, w, for the transformation matrix $\mathbf{T}$ is determined by rank($\mathbf{U}^{-1}\mathbf{V}$), where

$$\begin{aligned} \text{rank}(\mathbf{U}^{-1}\mathbf{V}) &\leq \text{rank}(\mathbf{V}) \\ &= \min(M, m \times (m-n) \times d/2) \\ &= \min(M, 9d \times (N^2-N)/2). \end{aligned} \tag{28}$$

The matrix $\mathbf{V}$ is full rank if $9d \times (N^2-N)/2 \geq M$; otherwise, the rank of matrix $\mathbf{V}$ is bounded by $9d \times (N^2-N)/2$, which mainly depends on the number of subjects N. Substituting M by $N \times n_i \times n$, the inequality $9d \times (N^2-N)/2 \geq M$ can be simplified to

$$N \geq 1 + \frac{2n_i}{3d}. \tag{29}$$

Given $n_i = d$ and $n=3$ in our study, we can infer from the above inequality that the number of classes $N \geq 2$, which implies the matrix $\mathbf{V}$ is full rank of M. Therefore, the number of bases for the KDT matrix $\mathbf{T}$, w, is bounded to the number of training images M.

4.2.4. Complexity analysis

The proposed KDT optimization algorithm requires $\mathcal{O}(M^3)$ time where M is the number of training images. As shown in Fig. 7, the computational complexity is discussed by splitting the proposed algorithm into four major operations, including singular value decomposition (SVD) (line 2, 9), matrix multiplication (line 5, 11, 13, 14), matrix inversion (line 10, 15) and eigenvalue decomposition (line 15).

First, the computational time for performing an SVD on K_{ii} (line 2) and $\mathbf{Q}_i^T\mathbf{Q}_j$ (line 9) are $\mathcal{O}(n_i^3)$ and $\mathcal{O}(d^3)$, respectively. Second, for each transformed kernel subspace, the complexity of matrix multiplication (line 5) and QR decomposition (line 6) is $\mathcal{O}(wd \times Mn_i)$ and $\mathcal{O}(w^2 \times d)$, respectively. In addition, the complexity of computing matrix $\mathbf{Z}$ (line 11) is $\mathcal{O}(M^2 \times dn_i)$, and those for within-scatter matrix $\mathbf{U}$ (line 13) and between-scatter matrix $\mathbf{V}$ (line 14) are $\mathcal{O}(M^2 \times dn_w)$ and $\mathcal{O}(M^2 \times dn_b)$, respectively, where n_w and n_b represent the number of within-subspaces and the number of between-subspaces. Since $\mathbf{R}_i$ (line 10) is $d \times d$ and $\mathbf{U}$ (line 15) is $M \times M$, we refer the complexities of computing $\mathbf{R}^{-1}$ and $\mathbf{U}^{-1}$ to $\mathcal{O}(d^2 \log d)$ and $\mathcal{O}(M^2 \log M)$, respectively. Finally, in line 15, the eigenvalue problem costs $\mathcal{O}(M^3)$ computations to obtain eigenvectors of $\mathbf{U}^{-1}\mathbf{V}$, where M is the number of columns (or rows) of $\mathbf{U}^{-1}\mathbf{V}$. Overall, the entire computational complexity is referred to $\mathcal{O}(M^3)$, where solving the eigenvalue problem in line 15 is the most significant part for the whole algorithm, since the number of training data M is often larger than n_i, n_w, n_b and the number of bases of kernel subspace d. The overall computational complexity is summarized in Table 1.

4.3. Reference subspace creation using the optimal KDT matrix

Above discussion is the acquisition of the optimal KDT matrix $\mathbf{T}$ that produces the maximal canonical differences of between-subspaces and the minimal canonical difference of within-subspaces. For all kernel subspaces $\mathbf{P}_i$, the orthonormal terms obtained from the QR-decomposition on each transformed kernel subspace $\mathbf{T}^T\mathbf{P}_i$ are defined as the reference subspace:

$$\text{Ref}_i = \mathbf{T}^T\mathbf{P}_i\mathbf{R}_i^{-1}. \tag{30}$$

Here, we have avoided the problem of performing operations of dot products in a high (or even infinite) dimensional feature space $\mathcal{F}$ by the use of kernel functions, i.e. each element of $\mathbf{T}^T\mathbf{P}_i$ is directly computable in accordance with Eq. (16).

Table 1
Summary of computational complexity of the proposed KDT method.

Operation	Details	Complexity
Singular value decomposition (SVD)	SVD on $\mathbf{K}_{ii}$	$\mathcal{O}(n_i^3)$
	SVD on $\mathbf{Q}_i^T\mathbf{Q}_j$	$\mathcal{O}(d^3)$
Matrix multiplication	Compute $\mathbf{T}^T\mathbf{P}_i$	$\mathcal{O}(wdMn_i)$
	QR-decomposition of $\mathbf{T}^T\mathbf{P}_i$	$\mathcal{O}(w^2d)$
	Compute matrix $\mathbf{Z}$	$\mathcal{O}(M^2dn_i)$
	Compute within-matrix $\mathbf{U}$	$\mathcal{O}(M^2dn_w)$
	Compute between-matrix $\mathbf{V}$	$\mathcal{O}(M^2dn_b)$
Matrix inversion	Inverse of $\mathbf{R}$	$\mathcal{O}(d^2\log d)$
	Inverse of $\mathbf{U}$	$\mathcal{O}(M^2\log M)$
Eigenvalue solution	Eigen-decomposition of $\mathbf{U}^{-1}\mathbf{V}$	$\mathcal{O}(M^3)$

Overall complexity: $\mathcal{O}(M^3)$.
n_i: the number of images per image set.
d: the number of bases of kernel subspace $\mathbf{P}_i$.
w: the number of bases of the KDT matrix $\mathbf{T}$.
M: the number of training images.

5. Test process

As shown in Fig. 2(b), the testing process begins by acquiring the image set $\mathbf{X}_{test}$ consisting of n_{test} 20×20-pixel facial images. The kernel subspace $\mathbf{P}_{test}$ corresponding to $\mathbf{X}_{test}$ is then generated by KPCA using Eqs. (3) and (4), and the transformed kernel subspace $\mathbf{T}^T\mathbf{P}_{test}$ is computed using the optimized KDT matrix $\mathbf{T}$ associated with Eq. (16). Eq. (30) is then applied to $\mathbf{T}^T\mathbf{P}_{test}$ to obtain the corresponding reference subspace Ref_{test}. Finally, by applying Eqs. (7) and (10) to the testing reference subspace and each of the training reference subspaces, respectively, a recognition result can be found as the class label c_i of the training reference subspace yielding the minimal canonical difference, i.e.

$$id = \underset{c_i}{\operatorname{argmin}} CanonicalDiff(i,test), \tag{31}$$

where id is the recognition result which belongs to the set of class labels, i.e. $id \in \{c_i | i=1,\ldots,m\}$ and $test$ represents the index of the testing image set $\mathbf{X}_{test}$.

6. Experimental results

We evaluate the performance of the proposed image set-based face recognition system using two databases, the Yale Face database B [45,46] and a self-compiled database. This section commences by describing the contents of the two databases and discussing the various categories of images within them. Experiments are then conducted to investigate the optimal parametric settings of the proposed KDT algorithm. Finally, our recognition performance are compared with existing still-image-based and image set-based methods.

6.1. Database contents

The Yale Face database B (shortened for convenience hereafter as YaleB) comprises 38 subjects and each subject contains 585 facial images with different head motions (or poses) and lighting conditions. Meanwhile, the self-compiled database consists of 32 subjects with arbitrary poses and either lighting variations or facial expression variations. The image sequences in the self-compiled database were recorded using a digital video camera at a rate of 30 fps with a resolution of 320×240 pixels. For both databases, we cropped out facial images from each frame using the cascaded face detection algorithm [47]. Each facial image was resized to the 20×20-pixel resolution to compare our performance with existing image set-based methods [22,31,32], where the authors also employed 15×15 or 20×20 resolution for experiments.

For the YaleB database, we selected facial images with the light source azimuth and elevation angle ranging from -85° to $+85^\circ$ and -40° to $+40^\circ$, respectively. A positive azimuth or panning angle implies that the light source direction or facial pose are toward the right side of the subject, while a negative angle implies that it is toward the left side. Similarly, a positive elevation or tilt angle implies that the light source direction or facial pose are above the horizon, while a negative angle implies the inverse direction. Each facial pose contains one of the nine simultaneous pan and tilt rotations according to inclinations of 0°, 12° or 24° from the optical axis of the camera [45,46]. For the self-compiled database, we collected the facial images associated with lighting direction of azimuthal angles varying from -90° to $+90^\circ$ and an elevation angle of 0°. Each facial image exhibited an arbitrary pose setting of a pan angle ranging from -45° to $+45^\circ$ and a tilt angle from -40° to $+40^\circ$. In addition, we compiled facial images associated with four common facial expressions including neutral, smile, surprise and disgust. These images were acquired under the same pose settings and indoor-ambient lighting conditions.

In order to investigate the recognition capabilities of the proposed system for facial images with variations such as lighting conditions and facial expressions, we sort the images of each database into three categories (denoted as *LightingYaleB*, *Lighting* and *Expression*) in accordance with their appearance variations. As summarized in Tables 2 and 3, the facial images of each subject within each category are manually partitioned into several groups to represent different varieties. For example, in the *LightingYaleB* category, 120 dissimilar facial images of each subject were by hand assigned to each of the three groups, *YaleB1*, *YaleB2* and *YaleB3*, according to the azimuthal position of the light source. Each subject therefore contains a total of 360 (3 groups × 120) facial images in this category. The same procedure of partitioning groups associated with their appearance variation was then applied to build up the *Lighting* and *Expression* categories. Note that each group of these categories contains 120 dissimilar facial images for each subject. Fig. 8 shows typical facial images within the *LightingYaleB*, *Lighting* and *Expression* categories. In the following experiments, we randomly separated the 120 facial images of each subject into four image sets to capture the scattering of within-subspaces.

6.2. Optimal KDT algorithm parametric settings

This section discusses the parametric settings by analyzing the convergence properties of the proposed KDT algorithm and investigating its sensitivity to the number of bases w of the KDT matrix $\mathbf{T}$. To reduce the variation in evaluating the parameters, the experiments were performed using images selected from one group in each category, namely *YaleB2*, *FrontalLight* and *Neutral*. In each group, four image sets, each of which contains 30 images, were selected for each subject: three for training purposes and one for testing. The variance term σ of the Gaussian RBF kernel in Eq. (2) was set as 0.05.

6.2.1. Convergence property of the KDT algorithm

The proposed KDT algorithm obtains an optimal KDT matrix $\mathbf{T}$ through an iterative learning procedure until the Jacobian value $J(\alpha)$ changes less than a small value ε. As shown in Fig. 7, the algorithm starts by a random initialization of the coefficient matrix α. In order to explore the convergence property of the

Table 2
Details of facial images in *LightingYaleB* category in YaleB database [45,46] and *Lighting* category in self-compiled database.

Category	Group	Light source direction	Facial pose
LightingYaleB (lighting+pose)	*YaleB1*	Azimuth: -85° to -30° Elevation: -40° to $+40^\circ$	Pan: -24°, -12°,0° Tilt : -24°,-12°,0°,$+12^\circ$,$+24^\circ$
	YaleB2	Azimuth: -30° to $+30^\circ$ Elevation: -40° to $+40^\circ$	
	YaleB3	Azimuth: $+30^\circ$ to $+85^\circ$ Elevation: -40° to $+40^\circ$	
Lighting (lighting+pose)	*LeftLight*	Azimuth: -90°, -60°, -30° Elevation: 0°	Pan: -45° to $+45^\circ$ Tilt: -40° to $+40^\circ$
	FrontalLight	Azimuth: -30°, 0°, $+30^\circ$ Elevation: 0°	
	RightLight	Azimuth: $+30^\circ$, $+60^\circ$, $+90^\circ$ Elevation: 0°	

Table 3
Details of facial images in *Expression* category in self-compiled database.

Category	Group	Facial expression	Facial pose
Expression (expression+pose)	*Neutral* *Smile* *Disgust* *Surprise*	Neutral Smile Disgust Surprise	Pan: -45° to $+45^\circ$ Tilt: -40° to $+40^\circ$

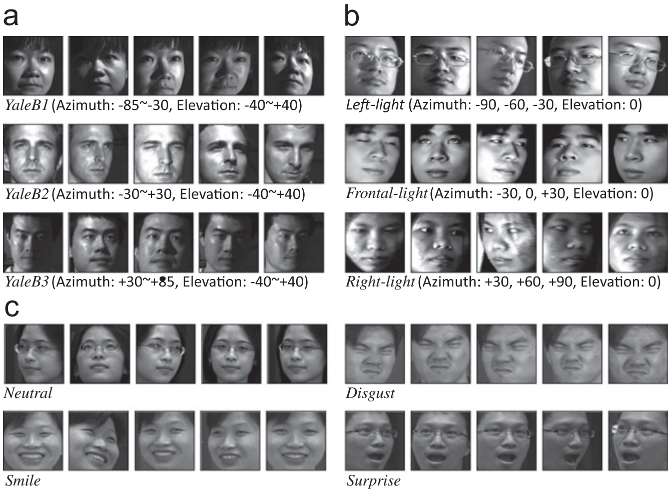
a
b
YaleB1 (Azimuth: -85~-30, Elevation: -40~+40)
Left-light (Azimuth: -90, -60, -30, Elevation: 0)
YaleB2 (Azimuth: -30~+30, Elevation: -40~+40)
Frontal-light (Azimuth: -30, 0, +30, Elevation: 0)
YaleB3 (Azimuth: +30~+85, Elevation: -40~+40)
Right-light (Azimuth: +30, +60, +90, Elevation: 0)
c
Neutral
Disgust
Smile
Surprise

Fig. 8. Typical examples of facial images in *LightingYaleB*, *Lighting* and *Expression* categories. (a) *YaleB1*, *YaleB2*, *YaleB3* groups (corresponding to different azimuthal angle ranges and a constant elevation angle range of $-40^\circ-+40^\circ$), (b) *LeftLight*, *FrontalLight*, *RightLight* groups (corresponding to different azimuthal angle ranges and a constant elevation angle of 0°), and (c) *Neutral*, *Smile*, *Disgust*, and *Surprise* groups.

KDT algorithm, the training process was performed using three randomized initializations of α with three selected image groups *YaleB2*, *FrontalLight* and *Neutral*. Figs. 9(a)–(c) illustrate the corresponding convergence curves for the three groups. The y-axis indicates $J(\alpha)$ computed in Eq. (22), while the x-axis indicates the number of iterations of the KDT algorithm. It can be seen that the convergence curves for *YaleB2* and *FrontalLight* oscillate more severely than that for *Neutral* because the training data in the former groups are affected more severely by lightings and distribute more complex for optimization. In every group, however, $J(\alpha)$ tends to converge toward a similar value with the increasing number of iterations regardless of the initial value assigned to α. In this experiment, the value of $J(\alpha)$ is apt to be stable within approximately 3.4 iterations on average. The result implies that the KDT algorithm would converge among a number of iterations using arbitrary initialization of α. Moreover, it is observed that the $J(\alpha)$ convergence curves are very similar in these figures, which could be inferred that the KDT algorithm provides a robust optimization performance against the nature of the appearance variations in the training images.

6.2.2. The number of bases for the KDT matrix

As shown in Eq. (15), the KDT matrix $\mathbf{T}$ is directly related to the total number of training images M. To determine an appropriate number of bases w for $\mathbf{T}$, a series of experiments were performed using different values of w with *YaleB2* and *Neutral* groups. For the *YaleB2* group, the training experiments were conducted using three training sets of different sizes: $N=12,22,32$ subjects corresponding to $m=36,66,96$ ($m=N\times n$, $n=3$) image sets and a total of $M=1080,1980,2880$ ($M=m\times n_i$, $n_i=30$) images, respectively. Similarly, for the *Neutral* group, the number of training sets $N=12,22,32$ in accordance with $m=48,88,128$ ($m=N\times n$, $n=4$) image sets and $M=1440,2640,3840$ ($M=m\times n_i$, $n_i=30$) training images.

We evaluate the KDT algorithm by comparing its performance with the KCMSM method which also considers a discriminative transformation subspace for all kernel subspaces in the feature space $\mathcal{F}$. Fig. 10 illustrates the relationship between the recognition rate of these methods and the dimensionality w in terms of the percentage of M. Figs. 10(a)–(c) present the results for the *YaleB2* group, while (d)–(f) present those for the *Neutral* group. Although the difference of the average performance between the upper and lower sets of figures is probably meaningless, the recognition rate is higher in *Neutral* than in *YaleB2*, which could be also explained by complicated manifolds of associated data groups. It can be seen that after the number of bases w for the KDT matrix reaches a certain value (here $w\approx 0.7M$), the identification rate of KDT saturates at an approximately constant value of 100%. However, at lower values of w, both KCMSM and KDT methods suffer from inferior accuracy due to the lack of the number of bases for the transformation subspace to provide sufficient discrimination. Apparently, the choice of w for the KDT matrix is more insensitive to the size of the training set M than that for KCMSM when w is large enough for discriminating these kernel subspace.

6.3. System performance evaluation

The recognition performance of the proposed system was evaluated by conducting experiments corresponding to two experimental protocols, denoted as *Experiment1* and *Experiment2*.

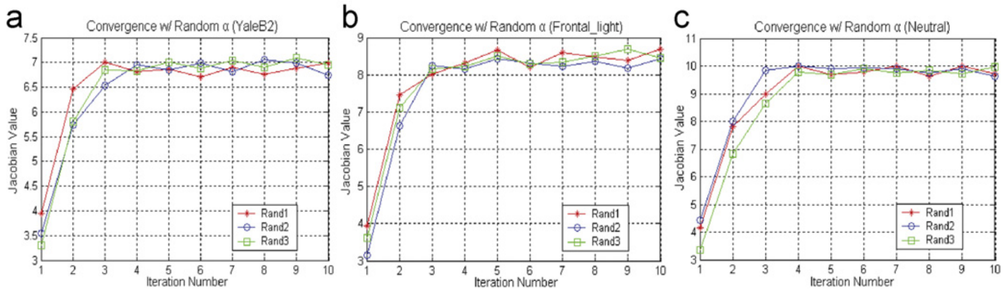

Fig. 9. Convergence of Jacobian function $J(\alpha)$ with respect to the number of iterations of KDT algorithm. Note that (a), (b) and (c) correspond to the *YaleB2*, *FrontalLighting* and *Neutral* groups, respectively, with three different initializations of α in every case.

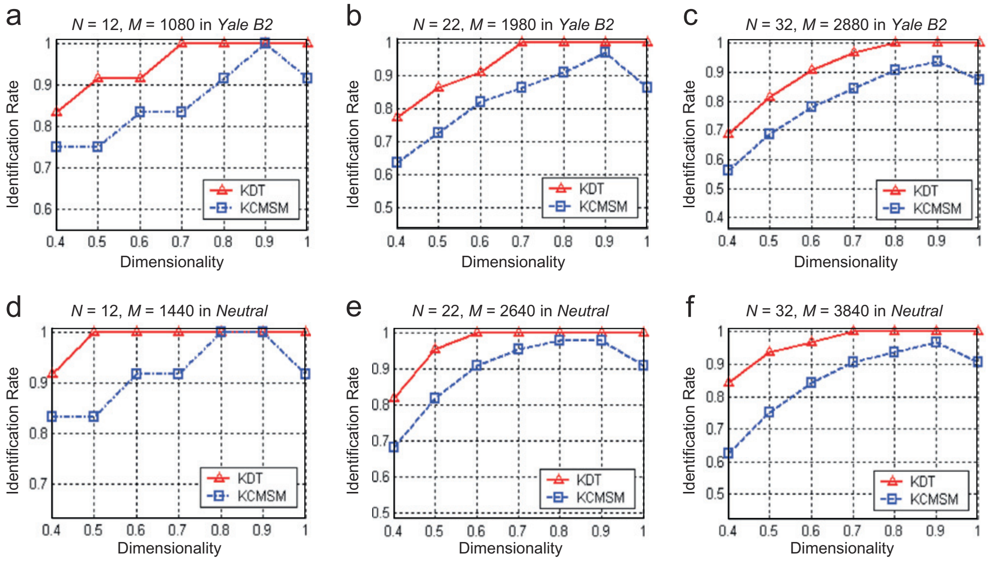

Fig. 10. Variation of identification rate with choice of dimensionality w of KDT matrix $\mathbf{T}$. (a)–(c) illustrate the identification rates for $N=12,22,32$ training subjects in the *YaleB2* group, respectively, corresponding to $m=36,66,96$ and $M=1080,1980,2880$ training images. (d)–(f) illustrate the identification rates for $N=12,22,32$ training subjects in the *Neutral* group, respectively, corresponding to $m=48,88,128$ and $M=1440,2640,3840$ training images.

In the first protocol, three image sets per subject were randomly selected from one group for training purposes, while the remaining image sets in the same group were used for testing. We conduct the experiments by using three training/testing combinations of image sets to report the recognition rate for each group.

In the second protocol *Experiment2*, one image set was randomly selected from each group in a category for training purposes, and the testing process uses two remaining image sets out of each group within the same category. As *Experiment1*, this protocol is also performed three times with different training/testing combinations. The difference between two protocols is that in *Experiment1*, the training and testing data were compiled by image sets from one group, while in *Experiment2*, images were selected from different groups, and therefore extends *Experiment1* to a wider-range of challenges. Specifically, we intend to evaluate the performance in a controlled environmental variation using *Experiment1* protocol and a wider environmental variation using *Experiment2*. For example, the azimuthal angle of the light source varies from -85° to -30° in *YaleB1* of *Experiment1*, but is extended to the range $-85^\circ-+85^\circ$ when conducting *Experiment2* using the facial images in the *LightingYaleB* category.

For consistent comparisons between different methods in the following experiments, we use $N=32$ subjects with each subject having $n=3$ training image sets in *LightingYaleB* and *Lighting* categories, and $n=4$ image sets in the *Expression* category. Based on the results discussed in Section 6.2, the number of iterations and the number of bases w is, respectively, fixed at 5 and $0.7M$ for the evaluation trails. A further series of experiments was performed using the *Experiment2* protocol to compare the recognition performance of the proposed system with those of two still-image-based as well as four image set-based methods.

6.3.1. Evaluation of system performance under different poses, lighting conditions and facial expressions

In developing automatic face recognition schemes for practical surveillance systems, robustness toward varying poses, light conditions and facial expression is particularly important since the testing images are possibly acquired at any location and any time of the day or night. We evaluate the robustness of the proposed system to these variations by applying *Experiment1* and *Experiment2* protocols to the *LightingYaleB*, *Lighting* and *Expression* categories. Each protocol was performed three times to ensure an unbiased comparison of performance between different groups.

As summarized in Table 4, the average recognition rate using the *Experiment1* protocol is 0.8 percentage points higher compared to those using *Experiment2*, because *Experiment1* includes the same range of appearance variation for both training and testing processes such that the KDT matrix provides a more discriminant transformation for testing subspaces. In the *LightingYaleB* category, the group of *YaleB2* obtains 1.2 percentage points better performance than the other two groups. The lighting

Table 4
Summary of face recognition results obtained using *Experiment1* and *Experiment2* protocols for facial images in *LightingYaleB*, *Lighting* and *Expression* categories. Note that for both protocols, each experiment is performed on three training/testing combinations of image sets for reporting the recognition rate.

Category	Group	*Experiment1* (%)	*Experiment2* (%)
LightingYaleB	*YaleB1*	97.7 ± 1.56	96.9 ± 1.56
	YaleB2	99.2 ± 1.56	97.9 ± 0.90
	YaleB3	98.4 ± 1.80	96.9 ± 0.90
Lighting	*LeftLight*	98.4 ± 1.80	97.4 ± 0.90
	FrontalLight	99.2 ± 1.56	98.4 ± 1.56
	RightLight	99.2 ± 1.56	98.4 ± 1.56
Expression	*Neutral*	100.0 ± 0.00	98.4 ± 0.00
	Smile	99.2 ± 1.56	98.4 ± 1.56
	Disgust	98.4 ± 1.80	97.9 ± 2.39
	Surprise	97.7 ± 1.56	98.4 ± 1.56
	Average	98.7	97.9

from the front of faces involves slighter influences on providing discriminatory subspace information for classification than that from the lateral of faces. It is not surprising that larger variations across data behaviors result in more loss of recognition rates. By similar explanation, the *FrontalLight* and the *Neutral* achieve 0.4 and 1.6 percentage points higher accuracy over the other groups in *Lighting* and *Expression* categories, respectively. The average recognition rate of *Lighting* exceeds that of *LightingYaleB* by 0.5 percentage points because the illumination setting of the YaleB database is much sharper. Overall, the proposed KDT algorithm correctly recognizes average 98.2% of the subjects using two protocols in these categories.

6.3.2. *Comparison of performance of the proposed system with existing still-image-based systems*

In this section, we compare the recognition performance of the proposed system with those of still-image-based methods. The experimental procedure was performed using *Experiment2* protocol described above. The performance of still-image-based methods was evaluated by *k*-nearest neighbor (*k* NN) of images transformed by PCA and LDA [9,26] subspaces. We chose the dimensionality of PCA subspace as 130 to preserve 98% of training data energy. Note that the first three eigenvectors were removed to reduce the lighting variations. The dimensionality of LDA subspace was set at 37 for the YaleB database and 31 for the self-compiled database. In the evaluation of *k* NN methods, we examined the *k* nearest projection vectors for each image set in the PCA or LDA subspaces. The number of bases of the KDT matrix **T** is specified as 0.7*M*, i.e. $w=2016$ for *LightingYaleB* and *Lighting* categories and $w=2688$ for the *Expression* category. Based on a voting process, the recognition result for each testing image set was obtained as the training set which contributes the most to these projection vectors.

The training image sets for both still-image-based methods and the KDT method are selected as the same, while two remaining image sets for each group are randomly selected to report the recognition result. The recognition results obtained by still-image-based methods and the KDT method within the three categories are summarized in Table 5, where each entry gives the average recognition rate and the corresponding standard deviation. Totally we utilized 576 testing image sets for *LightingYaleB* and *Lighting* categories and 768 for the *Expression* category. As shown, *k* NN-PCA gave the worst recognition rate among all methods in this experiment. On the other hand, *k* NN-LDA computes the discriminative subspace for all input images and yields a 7.0 percentage points improvement over *k* NN-PCA on average. The KDT method utilizes multiple testing images to span corresponding subspaces rather than testing each still image separately, and therefore contains more distinct information for classification. As a result, the proposed KDT method outperforms conventional still-image-based methods and obtains an average recognition rate of 97.9 percentage.

Table 5
Comparison of face recognition results (Avg ± Dev%) obtained using proposed KDT method and still-image-based methods in *LightingYaleB*, *Lighting* and *Expression* categories. The experimental procedures are performed in accordance with the *Experiment2* protocol.

Category	1NN-PCA (%)	1NN-LDA (%)	10NN-PCA (%)	10NN-LDA (%)	KDT (%)
LightingYaleB	82.3 ± 1.7	89.4 ± 1.4	85.7 ± 1.6	92.4 ± 1.0	97.2 ± 0.6
Lighting	83.5 ± 1.5	90.8 ± 1.2	86.8 ± 1.2	93.6 ± 0.6	98.1 ± 0.5
Expression	84.3 ± 1.4	91.2 ± 1.3	87.1 ± 1.3	94.3 ± 0.5	98.3 ± 0.3
Average	83.3	90.5	86.5	93.4	97.9

6.3.3. *Comparison of the proposed system with existing image set-based systems*

Five existing image set-based methods, namely mutual subspace method (MSM) [19], discriminant-analysis of canonical correlation (DCC) [22], the nonlinear kernel mutual subspace method (KMSM) [33,36], kernel orthogonal mutual subspace method (KOMSM) [32] and the proposed kernel discriminant transformation (KDT) method were compared. The parameters used in MSM, DCC, KMSM and KOMSM were also carefully tuned referring to original studies. In every case, the experimental procedure was employed using the *Experiment2* protocol described in Section 6.3.1. Totally 576 testing image sets was utilized for the *LightingYaleB* category and the *Lighting* category while 768 testing image for the *Expression* category. The settings for the KDT matrix **T** is the same as that mentioned in Section 6.3.2. Meanwhile, for MSM and DCC methods, PCA was performed to compute the orthonormal basis matrix for the linear subspace of each image set; the number of bases of each linear subspace was assigned to 8 to preserve 98 percentage of data energy. For kernel-based methods, i.e. KMSM, KOMSM and KDT, the number of bases of each kernel subspace was specified as 30 ($d=30$) while a value of $\sigma = 0.05$ was used for the Gaussian RBF kernel based on the result of preliminary experiments. The number of bases of the transformation matrix of the DCC method was specified as 350. The dimensionality of the whitening transformation matrix in the KOMSM method was also assigned to a value of 70% of the number of training images as the same number of bases *w* of the KDT matrix **T**.

The recognition results were obtained by applying the five methods to each group of the three categories. As summarized in Table 6, each entry gives the average recognition rate and the corresponding standard deviation over the same training/testing combinations as the experiments on still-image-based methods in Section 6.3.2. Despite that 10NN-LDA is a still-image-based method, it is interesting that it obtains similar performance as the MSM method. The reason might be that LDA transforms all inputs to a more discriminative subspace while MSM does not. Besides, MSM also lacks the ability of tackling nonlinear structures even though it achieves higher performance then 10NN-PCA by 6.3 percentage points on average. On the other hand, the KMSM method includes nonlinear manifold modeling by applying the "kernel trick" to MSM, and brings about 1.4 percentage points arise of the recognition rate. Again, analogous to MSM, KMSM faces the problem of lacking discriminative transformation and yields the worst performance of the three comparative kernel methods.

Table 6
Comparison of face recognition results (Avg ± Dev%) obtained using the proposed KDT method and four existing image set-based methods for *LightingYaleB*, *Lighting* and *Expression* categories.

Category	MSM (%)	KMSM (%)	DCC (%)	KOMSM (%)	KDT (%)
LightingYaleB	91.7 ± 1.3	93.4 ± 1.2	95.1 ± 0.8	95.3 ± 0.6	97.2 ± 0.6
Lighting	93.4 ± 1.1	94.4 ± 0.9	96.5 ± 0.5	96.5 ± 0.5	98.1 ± 0.5
Expression	93.5 ± 1.0	94.8 ± 1.0	96.9 ± 0.4	97.0 ± 0.4	98.3 ± 0.3
Average	92.8	94.2	96.1	96.3	97.9

The DCC method, considering a linear discriminant function that maximizes the canonical correlations of within-class image sets and minimizes those of between-class image set, obtains better performance over MSM and KMSM methods in terms of 3.3 and 1.9 percentage points of recognition rate. The KOMSM method provides greater accuracy by modifying KMSM with a pre-application of a kernel whitening transformation matrix. While the result of KOMSM is about 2.1 percentage points greater than KMSM, the proposed KDT algorithm is superior to all existing methods discussed in this literature. For the *LightingYaleB* category compiled by facial images under more severe lighting conditions, the recognition result is about 1.0 percentage point lower than that of the other two categories. For the self-compiled *Lighting* category, the KDT algorithm still contains better recognition performance compared to other existing methods using three data categories. It is notable that the lowest standard deviation of the recognition rate is obtained and confirms the robustness of the proposed system.

6.3.4. *Experiments on large dataset: Labeled Faces in the Wild*

The Labeled Faces in the Wild (LFW) [48] dataset offers a collection facial images captured from the media. It consists of 1680 subjects with two or more images. While our method is intended for recognizing facial images based on image sets rather than single images, we ignore the official LFW experimental protocol and devise our own. Thus, we use a subset of the LFW dataset which consists of 100 subjects having at least 10 images. For this dataset, each facial image was cropped using the cascade face detector [47], and resized into 20 × 20 resolution. Fig. 11 shows example images from the LFW dataset used in our experiments. Note that the facial images consist of more uncontrolled variations such as head poses, facial expression, and lighting conditions.

Since LFW has too few images per person for our framework, we generate three more variations for each facial image. In the end we have 40 facial images for each subject. The 40 images were randomly separated into four image sets. Because LFW do not have specific pose and illumination settings as the YaleB and the self-compiled dataset, *Experiment2* protocol is not applicable in the case. Thus, we use *Experiment1* protocol to conduct the performance evaluation across the above image set-based methods. Note that we performed *Experiment1* three times, and reported the results using 1200 testing image sets. The results are summarized in Table 7.

While the LFW contains too less images for image set-based face recognition framework, we show, again, that MSM and KMSM perform as the baseline methods as in the previous experiments, since they do not consider discriminative information between subspaces. Also, the LFW facial images contain more uncontrolled variations than YaleB and the self-compiled dataset, thus the standard variation of recognition accuracy is larger in this experiment. From this point of view, kernel-based approaches, i.e. KOMSM and KDT, perform 2.3 and 5.6 percentage points arise than DCC by considering more nonlinearity, which better describes the distribution of facial image sets.

Fig. 11. Examples of facial images in the LFW dataset [48]. Different rows show different persons, respectively.

Table 7
Comparison of face recognition results (Avg ± Dev%) on the LFW dataset.

Category	MSM (%)	KMSM (%)	DCC (%)	KOMSM (%)	KDT (%)
LFW	30.7 ± 5.9	32.3 ± 4.0	59.7 ± 3.1	62.0 ± 4.4	65.3 ± 3.2

6.3.5. *Statistical significance analysis*

We performed standard hypothesis testing techniques [49] to verify the statistical significance in the comparison of five image set-based methods listed in Tables 6 and 7. When comparing two methods $\mathcal{A}$ and $\mathcal{B}$, our hypothesis and the null hypothesis are given by

$\mathbb{H}0$ $\mathcal{A}$ correctly recognizes testing image sets more often than $\mathcal{B}$.
$\mathbb{H}1$ There is no difference how well $\mathcal{A}$ and $\mathcal{B}$ perform.

The probability of $\mathbb{H}0$ is determined by a normalized variable z according to the number of times that each method is succeed to recognize the testing image set. The test statistic z is given by

$$z = \frac{p_{\mathcal{A}} - p_{\mathcal{B}}}{\sqrt{\frac{2P_c(1-p_c)}{m_{test}}}}, \tag{32}$$

where p_A and p_B are the observed proportions of success in the testing image sets for method $\mathcal{A}$ and $\mathcal{B}$, $p_c = (p_{\mathcal{A}} + p_{\mathcal{B}})/2$, and m_{test} is the number of testing image sets. Table 8 shows the results of significant testing in the comparison of the KDT method and other image set-based methods, including the number of image sets correctly recognized by each method ($m_{correct}$), z value, and probability of $\mathbb{H}0$ ($P_{\mathbb{H}0}$). We used a common cutoff level at 0.05, i.e. $\mathbb{H}0$ is rejected if $P_{\mathbb{H}0} \leq 0.05$. Using this cutoff level, we analyze the statistical significance between each pair of KDT and the others.

We observe that for each pairwise comparison, KDT is demonstrated to be statistically significant with a 0.05 cutoff. We can also observe from the comparison of KDT between MSM and KMSM that there are highly statistical significance with $P_{\mathbb{H}0} \leq 0.002$. This means that the possibility of $\mathbb{H}0$ being correct is less than or equal to 0.002. As a result, the KDT algorithm was demonstrated to have statistically significant differences among five image set-based methods discussed in this literature.

Table 8
Results of statistical significance testing in the comparison of the proposed KDT method and four image set-based methods. Each testing image set are treated as independent samples for each method. There are totally 576 testing image sets for *LightingYaleB* and *Lighting* categories, 768 for the *Expression* category, and 1200 for the *LFW* dataset. Note that m_{correct} stands for the number of image sets correctly recognized by each method; in this table, the left and right column of m_{correct} corresponds to the left and right column of Methods, respectively.

Category	Methods		m_{correct}		Variable z	$P_{\mathbb{H}_0} \le$
LightingYaleB	MSM	KDT	528	560	4.076	0.00002
	KMSM	KDT	537	560	3.047	0.00116
	DCC	KDT	548	560	1.852	0.03201
	KOMSM	KDT	549	560	1.697	0.04485
Lighting	MSM	KDT	539	566	4.061	0.00002
	KMSM	KDT	544	566	3.416	0.00032
	DCC	KDT	556	566	1.796	0.03625
	KOMSM	KDT	556	566	1.796	0.03625
Expression	MSM	KDT	718	755	4.743	0.00000
	KMSM	KDT	728	755	3.758	0.00009
	DCC	KDT	744	755	1.792	0.03657
	KOMSM	KDT	745	755	1.681	0.04638
LFW dataset	MSM	KDT	368	784	16.997	0.00000
	KMSM	KDT	388	784	16.171	0.00000
	DCC	KDT	716	784	2.867	0.00207
	KOMSM	KDT	744	784	1.698	0.04470

7. Conclusions

This study introduces an image set-based face recognition system based upon a novel kernel discriminant transformation (KDT) algorithm. While the KDT matrix **T** cannot be explicitly calculated, the proposed algorithm reformulates the original problem as the kernel Fisher's discriminant to implicitly optimize **T** using an iterative procedure. We have shown that the number of bases for **T** is bound to the number of training images M and the computational complexity of the algorithm is $\mathcal{O}(M^3)$. Because the proposed system utilizes image sets rather than single testing images, the KDT algorithm increases the discriminant information as well as improves the robustness toward variations in poses, facial expressions and lighting conditions of the facial images. The experimental results have shown that the KDT optimization algorithm converges irrespective of the initialization conditions. The results have also shown that the optimal number of bases for **T** is basically insensitive to the number of images used in the training dataset. Finally, the proposed face recognition system has shown to yield a better recognition performance than those of existing still-image-based and image set-based methods presented in the literature.

For the case of a single testing image, the KDT algorithm proposed in this study is also applicable by considering the single image as a point in high-dimensional space, and then projecting this point onto corresponding kernel subspaces to measure its similarity. The computational complexity of the KDT algorithm increases significantly with an increasing number of training images. Therefore, future studies will investigate the feasibility of using an ensemble learning technique to reduce the number of training images required whilst preserving the quality of the classification results. In addition, the use of sequential learning methods such as incremental principal component analysis [49] and incremental linear discriminant analysis [50] will be considered as a means of improving the computational efficiency of the KDT optimization algorithm.

Acknowledgment

This work was supported by the National Science Council, Republic of China, Taiwan, under Contract NSC-99-2221-E-006-216.

References

[1] P.C. Yuen, G.C. Feng, D.Q. Dai, Human face image retrieval system for large database, International Conference on Pattern Recognition, vol. 2, 1998.

[2] A. Martinez, Face image retrieval using HMMs, in: IEEE Workshop on Content-Based Access of Image and Video Libraries, 1999, pp. 35–39.

[3] R. Chellappa, C.L. Wilson, S. Sirohey, et al., Human and machine recognition of faces: a survey, Proceedings of the IEEE 83 (5) (1995) 705–740.

[4] A. Samal, P.A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition 25 (1) (1992) 65–77.

[5] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys (CSUR) 35 (4) (2003) 399–458.

[6] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, et al., Eigenfaces vs fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[7] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2) (2001) 228–233.

[8] P.S. Penev, J.J. Atick, Local feature analysis: a general statistical theory for object representation, Network: Computation in Neural Systems 7 (3) (1996) 477–500.

[9] M. Turk, A. Pentland, Face recognition using eigenfaces, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 591, 1991.

[10] X. Lu, Y. Wang, A.K. Jain, Combining classifiers for face recognition, IEEE International Conference on Multimedia and Expo, vol. 3, 2003, pp. 13–16.

[11] H. Kang, T. Cootes, C. Taylor, A comparison of face verification algorithms using appearance models, British Machine Vision Conference, vol. 2, 2002, pp. 477–486.

[12] X. Zhang, Y. Gao, Face recognition across pose: a review, Pattern Recognition 42 (11) (2009) 2876–2896.

[13] S.I. Choi, C. Kim, C.H. Choi, Shadow compensation in 2D images for face recognition, Pattern Recognition 40 (7) (2007) 2118–2125.

[14] A. Shashua, T. Riklin-Raviv, The quotient image: class-based re-rendering and recognition with varying illuminations, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2) (2001) 129–139.

[15] S.W. Lee, S.H. Moon, S.W. Lee, Face recognition under arbitrary illumination using illuminated exemplars, Pattern Recognition 40 (5) (2007) 1605–1620.

[16] T. Zhang, B. Fang, Y. Yuan, Y. Yan Tang, Z. Shang, D. Li, F. Lang, Multiscale facial structure representation for face recognition under varying illumination, Pattern Recognition 42 (2) (2009) 251–258.

[17] W.-C. Kao, M.-C. Hsu, Y.-Y. Yang, Local contrast enhancement and adaptive feature extraction for illumination-invariant face recognition, Pattern Recognition 43 (5) (2010) 1736–1747.

[18] M. Vasilescu, D. Terzopoulos, Multilinear independent components analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, p. 547.

[19] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: IEEE Conference on Automatic Face and Gesture Recognition, 1998, pp. 318–323.

[20] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, International Symposium of Robotics Research, vol. 12, 2003.

[21] T.-K. Kim, O. Arandjelović, R. Cipolla, Boosted manifold principal angles for image set-based recognition, Pattern Recognition 40 (9) (2007) 2475–2484.

[22] T.-K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1005.

[23] W.-S. Chu, C.-R. Huang, C.-S. Chen, Identifying gender from unaligned facial images by set classification, in: International Conference on Pattern Recognition, 2010, pp. 2636–2639.

[24] W.-S. Chu, J.-C. Chen, J.-J. Lien, Kernel discriminant analysis based on canonical differences for face recognition in image sets, Asian Conference on Computer Vision 4844 (2007) 700–711.

[25] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (5) (1998) 1299–1319.

[26] A. Hadid, M. Pietikainen, From still image to video-based face recognition: an experimental analysis, in: IEEE Conference on Automatic Face and Gesture Recognition, 2004, pp. 17–19.

[27] K.C. Lee, J. Ho, M.-H. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2003.

[28] X. Liu, T. Chen, Video-based face recognition using adaptive hidden Markov models, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2003.

[29] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (1–2) (2003) 214–245.

[30] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, IEEE Conference on Computer Vison and Pattern Recognition, vol. 1, 2005, p. 581.

[31] K. Fukui, B. Stenger, O. Yamaguchi, A framework for 3D object recognition using the kernel constrained mutual subspace method, Asian Conference on Computer Vision 3852 (2006) 315.

[32] K. Fukui, O. Yamaguchi, The kernel orthogonal mutual subspace method and its application to 3D object recognition, Asian Conference on Computer Vision 4844 (2007) 467.

[33] H. Sakano, N. Mukawa, Kernel mutual subspace method for robust facial image recognition, International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, vol. 1, 2000.
[34] G. Shakhnarovich, J.W. Fisher, T. Darrell, Face recognition from long-term observations, Lecture Notes in Computer Science (2002) 851–868.
[35] L. Wang, X. Wang, J. Feng, Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition, Pattern Recognition 39 (3) (2006) 456–464.
[36] L. Wolf, A. Shashua, Kernel principal angles for classification machines with applications to image sequence interpretation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2003.
[37] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.
[38] G. Dai, Y.T. Qian, Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition, in: IEEE International Conference on Image Processing, 2004, pp. 2697–2700.
[39] G. Dai, D.Y. Yeung, Y.T. Qian, Face recognition using a kernel fractional-step discriminant analysis algorithm, Pattern Recognition 40 (1) (2007) 229–243.
[40] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K.R. Muller, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing (1999) 41–48.
[41] M.-H. Yang, Kernel eigenfaces vs Kernel fisherfaces: face recognition using kernel methods, in: IEEE Conference on Automatic Face and Gesture Recognition, 2002, p. 215.
[42] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 230.
[43] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, J. Wang, An efficient kernel discriminant analysis method, Pattern Recognition 38 (10) (2005) 1788–1790.
[44] Y. Grandvalet, S. Canu, Adaptive scaling for feature selection in SVMs, Advances in Neural Information Processing Systems (2003) 569–576.
[45] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 643–660.
[46] K.C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.
[47] P. Viola, M.J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[48] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
[49] H. Zhao, P.C. Yuen, J.T. Kwok, A novel incremental principal component analysis and its application for face recognition, IEEE Transactions on Systems, Man, and Cybernetics, Part B 36 (4) (2006) 873–886.
[50] H. Zhao, P.C. Yuen, Incremental linear discriminant analysis for face recognition, IEEE Transactions on Systems, Man, and Cybernetics, Part B 38 (1) (2008) 210–221.

Wen-Sheng Chu is a full-time research assistant in the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA. He received his B.S. and M.S. degrees in computer science and information engineering from National Cheng Kung University in 2005 and 2007, respectively. His research interests mainly focus on computer vision and pattern recognition problems, especially those related to automatic face recognition, image retrieval, gender classification and common pattern discovery.

Ju-Chin Chen received her B.S., M. S. and Ph.D. degrees in Computer Science and Information Engineering from National Cheng Kung University, Tainan, Taiwan, in 2002, 2004 and 2010, respectively. She is now an assistant professor in the Department of Computer Science and Information Engineering at National Kaohsiung University of Applied Science, Taiwan. Her research interests lie in the fields of machine learning, computer vision and pattern recognition.

Jenn-Jier James Lien (M'00) received his M.S. and Ph.D. degrees in electrical engineering from Washington University, St. Louis, MO, and the University of Pittsburgh, Pittsburgh, PA, in 1993 and 1998, respectively. From 1995 to 1998, he was a research assistant at the Vision Autonomous Systems Center in the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA. From 1999 to 2002, he was a senior research scientist at L1-Identity (formerly Visionics) and a project lead for the DARPA surveillance project. He is now an associate professor in the department of computer science and information engineering at National Cheng Kung University, Taiwan.