Inter-modality Face Recognition

Dahua Lin[1] and Xiaoou Tang[1,2]

[1] Dept. of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
dhlin4@ie.cuhk.edu.hk
[2] Microsoft Research Asia, Beijing, China
xitang@microsoft.com

Abstract. Recently, the wide deployment of practical face recognition systems gives rise to the emergence of the inter-modality face recognition problem. In this problem, the face images in the database and the query images captured on spot are acquired under quite different conditions or even using different equipments. Conventional approaches either treat the samples in a uniform model or introduce an intermediate conversion stage, both of which would lead to severe performance degradation due to the great discrepancies between different modalities. In this paper, we propose a novel algorithm called Common Discriminant Feature Extraction specially tailored to the inter-modality problem. In the algorithm, two transforms are simultaneously learned to transform the samples in both modalities respectively to the common feature space. We formulate the learning objective by incorporating both the empirical discriminative power and the local smoothness of the feature transformation. By explicitly controlling the model complexity through the smoothness constraint, we can effectively reduce the risk of overfitting and enhance the generalization capability. Furthermore, to cope with the nongaussian distribution and diverse variations in the sample space, we develop two nonlinear extensions of the algorithm: one is based on kernelization, while the other is a multi-mode framework. These extensions substantially improve the recognition performance in complex situation. Extensive experiments are conducted to test our algorithms in two application scenarios: optical image-infrared image recognition and photo-sketch recognition. Our algorithms show excellent performance in the experiments.

1 Introduction

The past decade has witnessed a rapid progress of face recognition techniques and development of automatic face recognition (AFR) systems. In many of the face recognition systems, we are in confront of a new situation: due to the limitations of practical conditions, the query face images captured on spot and the reference images stored in the database are acquired through quite different processes under different conditions. Here we give two cases arising from practical demands. The first case is a surveillance system operating from morning to night in an adverse outdoor environment. To combat the weak illumination in the nights or cloudy days, the system employ infrared cameras for imaging and

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 13–26, 2006.

compare the infrared images with the optical images registered in the database, as shown in fig.1. In another case, the police call for a photo-sketch recognition system to recognize the identity of a suspect from a sketch when his photos are unavailable, as shown in fig.2. The images acquired by different processes, which we say are in different *modalities*, often present great discrepancies, thus it is infeasible to use a single model to carry out the comparison between these images. These new applications bring forward a great challenge to the face recognition systems and require new techniques specially designed for the *Inter-Modality Face Recognition*.

Before introducing our approach to the problem, we give a brief review on the statistical pattern recognition methods. An important difficulty for face recognition lies in the high dimension of the sample space. To alleviate the curse of the dimensionality, it is crucial to reduce the dimension while preserving the important information for classification. LDA (Fisherface)[1] is the most popular dimension reduction method for face recognition, which pursues a feature subspace to maximize the trace-ratio of the between-class scattering matrix and the within-class scattering matrix. To solve the singularity of within-class scatter matrix incurred by small sample size problem, a variety of improved LDA-based algorithms are proposed[2][3][4][5][6][7]. However, these algorithms fail to address the overfitting fundamentally. We argue that the poor generalization of LDA in the small sample size case originates from the formulation of the objective, which merely emphasize the separability of the training samples without considering the factors affecting the generalization risk.

In this paper, we propose a general algorithm for various inter-modality face recognition problems, where two issues arise: **1)** How to enable the comparison between samples in different modalities without the intermediate conversion? **2)** How to enhance the generalization capability of the model? To tackle the former issue, we propose a novel algorithm called *Common Discriminant Feature Extraction* as illustrated in fig.3, where two different transforms are simultaneously learned to transform the samples in both the query modality and the reference modality to a common feature space, where the discriminant features for the two modalities are well aligned so that the comparison between them is feasible. Motivated by the statistical learning theory[10] which states that the

Fig. 1. The optical images vs. the infrared images

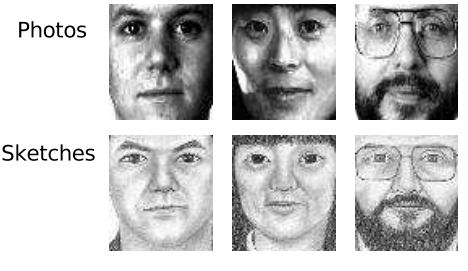

Fig. 2. The photos vs. the sketches

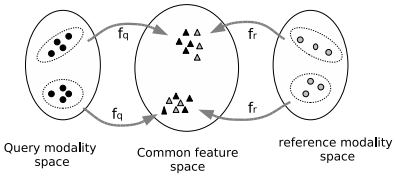

Fig. 3. Illustration of common feature space

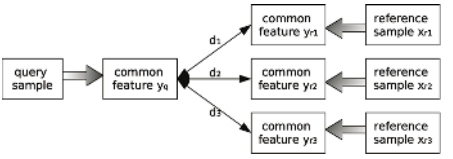

Fig. 4. The query procedure

model complexity has important impact on the generalization risk, we formulate the learning objective by incorporating both the empirical discriminative power and the local consistency. The empirical discriminative power comprises the intra-class compactness and inter-class dispersion, which together reflect the separability of the training samples; while the local consistency[11] is inspired by the local preservation principle emerging from the machine learning literatures[12][13], which measures the local smoothness of the feature transformation. It is believed that by explicitly imposing the smoothness constraint and thus preserving the local structure of the embedded manifold, we can effectively reduce the risk of overfitting. Based on the formulation, we derive a new algorithm which can efficiently solve the global optima of the objective function by eigen-decomposition.

Considering that linear transforms lack of capability to separate the samples well in the complicated situations where the sample distribution is nongaussian, we further derive two nonlinear extensions of the algorithm to exploit the nonlinearity of the sample space. The first extension is by kernelization, which offers an elegant and efficient way to extract nonlinear features. The second extension is a multi-mode framework. The framework learns multiple models adapting to the query samples captured in distinct conditions and makes the final decision by a belief-based weighted fusion scheme. Comprehensive experiments are conducted to validate the effectiveness of our algorithms.

2 Common Discriminant Feature Extraction

2.1 Formulation of the Learning Problem

In the problem, there are two types of samples: the query samples captured on spot and the reference samples stored in the database, which are in different modalities. The vector space of the query samples and the reference samples are denoted by $\mathcal{X}_q$ and $\mathcal{X}_r$ respectively, whose dimensions are denoted by d_q and d_r. Suppose we have a training set of N_q samples in the query space and N_r samples in the reference space from C classes, denoted by $\{(\mathbf{x}_i^{(q)}, c_i^{(q)})\}_{i=1}^{N_q}$ and $\{(\mathbf{x}_j^{(r)}, c_j^{(r)})\}_{j=1}^{N_r}$. Here $c_i^{(q)}$ and $c_j^{(r)}$ respectively indicates the class label of the corresponding sample. To enable the comparison of the query samples and the reference samples, we transform them to a d_c-dimensional *Common Discriminant Feature Space*, denoted by $\mathcal{Y}$, which preserves the important discriminant information and aligns the samples in two different modality so that the comparison

is feasible. We denote the transform for the query modality by $f_q : \mathcal{X}_q \to \mathcal{Y}$ and the transform for the reference modality by $f_r : \mathcal{X}_r \to \mathcal{Y}$. For succinctness of discussion, we denote $\mathbf{y}_i^{(q)} = f_q(\mathbf{x}_i^{(q)}; \theta_q)$ and $\mathbf{y}_j^{(r)} = f_r(\mathbf{x}_j^{(r)}; \theta_r)$, where θ_q and θ_r are the transform parameters. After the common feature space is learnt, the dissimilarity can be evaluated by transforming both the query sample and the reference sample to the common space and computing the distance between the feature vectors, as in fig.4.

To obtain the feature transforms with good generalization capability, we formulate the learning objective integrating both the empirical separability and the local consistency of the transform operators.

The empirical separability. The *empirical separability* describes the separability of the training samples. It involves two related goals: the intra-class compactness and the inter-class dispersion, which are measured by *average intra-class scattering* and *average inter-class scattering* respectively as follows:

$$J_1(\theta_q, \theta_r) = \frac{1}{N_1} \sum_{i=1}^{N_q} \sum_{j:c_j^{(r)}=c_i^{(q)}} ||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(r)}||^2, \tag{1}$$

$$J_2(\theta_q, \theta_r) = \frac{1}{N_2} \sum_{i=1}^{N_q} \sum_{j:c_j^{(r)}\neq c_i^{(q)}} ||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(r)}||^2, \tag{2}$$

where N_1 is the number of pairs of samples from the same class, N_2 is the number of pairs of samples from different classes. To better distinguish the samples from different classes, we should drive the query samples towards the reference samples from the same class and far from those of distinct classes. Based on the rationale, we derive the formulation of empirical separability by unifying the intra-class compactness and the inter-class dispersion:

$$J_e(\theta_q, \theta_r) = J_1(f_q, f_r) - \alpha J_2(f_q, f_r) = \sum_{i=1}^{N_q} \sum_{j=1}^{N_r} u_{ij} ||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(r)}||^2, \tag{3}$$

where $u_{ij} = \begin{cases} \frac{1}{N_1} & (c_i^{(q)} = c_j^{(r)}) \\ -\frac{\alpha}{N_2} & (c_i^{(q)} \neq c_j^{(r)}) \end{cases}$, and the α reflects the trade-off between the two goals. Minimization of $J_e(\theta_q, \theta_r)$ will lead to the feature space best separating the training samples.

The local consistency. To reduce the risk of overfitting, we introduce the notion *local consistency* into the formulation to regularize the empirical objective, which is a notion emerging from spectral learning[11] and manifold learning [14][12]. The local consistency for f_q and f_r are respectively defined by

$$J_l^{(q)}(\theta_q) = \frac{1}{N_q}\sum_{i=1}^{N_q}\sum_{j=1}^{N_q} w_{ij}^{(q)}||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(q)}||^2; \tag{4}$$

$$J_l^{(r)}(\theta_r) = \frac{1}{N_r}\sum_{i=1}^{N_r}\sum_{j=1}^{N_r} w_{ij}^{(r)}||\mathbf{y}_i^{(r)} - \mathbf{y}_j^{(r)}||^2, \tag{5}$$

where $\mathcal{N}(i)$ is the set of indices of the neighboring samples of i, $w_{ij}^{(q)} = \exp(-\frac{||\mathbf{x}_i^{(q)}-\mathbf{x}_j^{(q)}||}{\sigma_q^2})$ and $w_{ij}^{(r)} = \exp(-\frac{||\mathbf{x}_i^{(r)}-\mathbf{x}_j^{(r)}||}{\sigma_r^2})$ reflect the affinity of two samples. It has been shown that[14] such a definition corresponds to the approximation of $\int_{\mathcal{M}} ||\nabla f(\mathbf{x})||^2$ over the manifold $\mathcal{M}$ on which the samples reside. This clearly indicates that minimization of J_l will encourage consistent output for the neighboring samples in the input space, and thus result in the transform with high local smoothness and best locality preservation. Hence, a smooth transform that is expected to be less vulnerable to overfitting can be learnt by imposing the local consistency constraint.

Integrating the empirical objective and the local consistency objective, we formulate the learning objective to minimize the following objective function:

$$\begin{aligned} J(\theta_q,\theta_r) &= J_e(\theta_q,\theta_r) + \beta\left(J_l^{(q)}(\theta_q) + J_l^{(r)}(\theta_r)\right) = \sum_{i=1}^{N_q}\sum_{j=1}^{N_r} u_{ij}||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(r)}||^2 \\ &+ \sum_{i=1}^{N_q}\sum_{j=1}^{N_q} v_{ij}^{(q)}||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(q)}||^2 + \sum_{i=1}^{N_r}\sum_{j=1}^{N_r} v_{ij}^{(r)}||\mathbf{y}_i^{(r)} - \mathbf{y}_j^{(r)}||^2, \end{aligned} \tag{6}$$

where we introduce $v_{ij}^{(q)} = \frac{\beta w_{ij}^{(q)}}{N_q}$ and $v_{ij}^{(r)} = \frac{\beta w_{ij}^{(r)}}{N_r}$. For convenience. β is a regularization coefficient controlling the trade-off between the two objectives.

2.2 Matrix-Form of the Objective

To simplify the further analysis, we introduce the following matrix notations:

$d_c \times N_q$ matrix $\mathbf{Y}_q = \left[\mathbf{y}_1^{(q)}, \mathbf{y}_2^{(q)}, \ldots, \mathbf{y}_{N_q}^{(q)}\right]$, $d_c \times N_r$ matrix $\mathbf{Y}_r = \left[\mathbf{y}_1^{(r)}, \mathbf{y}_2^{(r)}, \ldots, \mathbf{y}_{N_r}^{(r)}\right]$
$N_q \times N_r$ matrix $\mathbf{U} : \mathbf{U}(i,j) = u_{ij}$,
$N_q \times N_q$ diagonal matrix $\mathbf{S}_q : \mathbf{S}_q(i,i) = \sum_{j=1}^{N_r} u_{ij}$, $N_r \times N_r$ diagonal matrix $\mathbf{S}_r : \mathbf{S}_r(j,j) = \sum_{i=1}^{N_q} u_{ij}$,
$N_q \times N_q$ matrix $\mathbf{V}_q : \mathbf{V}_q(i,j) = v_{ij}^{(q)}$, $N_r \times N_r$ matrix $\mathbf{V}_r : \mathbf{V}_r(i,j) = v_{ij}^{(r)}$,
$N_q \times N_q$ diagonal matrix $\mathbf{D}_q : \mathbf{D}_q(i,i) = \sum_{j=1}^{N_q} v_{ij}^{(q)}$, $N_r \times N_r$ diagonal matrix $\mathbf{D}_r : \mathbf{D}_r(i,i) = \sum_{j=1}^{N_r} v_{ij}^{(r)}$.

Then we can rewrite the objectives in matrix form as:

$$J_e(\theta_q,\theta_r) = \sum_{i=1}^{N_q}\sum_{j=1}^{N_r} u_{ij}||\mathbf{y}_i^{(q)} - \mathbf{y}_j^{(r)}||^2 = \mathrm{tr}\left(\mathbf{Y}_q\mathbf{S}_q\mathbf{Y}_q^T + \mathbf{Y}_r\mathbf{S}_r\mathbf{Y}_r^T - 2\mathbf{Y}_q\mathbf{U}\mathbf{Y}_r^T\right). \tag{7}$$

$$J_l^{(q)}(\theta_q) = 2\mathrm{tr}\left(\mathbf{Y}_q(\mathbf{D}_q - \mathbf{V}_q)\mathbf{Y}_q^T\right); \tag{8}$$

$$J_l^{(q)}(\theta_r) = 2\mathrm{tr}\left(\mathbf{Y}_r(\mathbf{D}_r - \mathbf{V}_r)\mathbf{Y}_r^T\right). \tag{9}$$

Combine the three formulas above, we can derive that

$$J(\theta_q, \theta_r) = \mathrm{tr}\left(\mathbf{Y}_q \mathbf{R}_q \mathbf{Y}_q^T + \mathbf{Y}_r \mathbf{R}_r \mathbf{Y}_r^T - 2\mathbf{Y}_q \mathbf{U} \mathbf{Y}_r^T\right). \tag{10}$$

where $\mathbf{R}_q = \mathbf{S}_q + 2(\mathbf{D}_q - \mathbf{V}_q)$ and $\mathbf{R}_r = \mathbf{S}_r + 2(\mathbf{D}_r - \mathbf{V}_r)$.

It is conspicuous that the transform $f(\mathbf{x})$ and its double-scaled version $2f(\mathbf{x})$ are essentially the same with respect to classification, however the latter transform will result in the objective value four times the former one. Hence, we should impose constraint on the scale of features in order to prevent trivial solutions. Since Euclidean distance will be used in the target feature space where all dimensions are uniformly treated, it is reasonable to require the feature vectors satisfy isotropic distribution. It can be expressed in terms of unit covariance as follows

$$\frac{1}{N_q} \mathbf{Y}_q \mathbf{Y}_q^T + \frac{1}{N_r} \mathbf{Y}_r \mathbf{Y}_r^T = \mathbf{I}. \tag{11}$$

2.3 Solving the Linear Transforms

Linear features are widely used in the literatures due to its simplicity and good generalization. Accordingly we first investigate the case where f_q and f_r are linear transforms, parameterized by the transform matrix $\mathbf{A}_q$ and $\mathbf{A}_r$. Denote the sample matrices[1] by $\mathbf{X}_q = \left[\mathbf{x}_1^{(q)}, \mathbf{x}_2^{(q)}, \ldots, \mathbf{x}_{N_q}^{(q)}\right]$ and $\mathbf{X}_r = \left[\mathbf{x}_1^{(r)}, \mathbf{x}_2^{(r)}, \ldots, \mathbf{x}_{N_r}^{(r)}\right]$, then we have

$$\mathbf{Y}_q = \mathbf{A}_q^T \mathbf{X}_q \qquad \mathbf{Y}_r = \mathbf{A}_r^T \mathbf{X}_r \tag{12}$$

Combining Eq.(10), Eq.(11) and Eq.(12), the optimization problem of the transform matrices $\mathbf{A}_q$ and $\mathbf{A}_r$ is given by

$$\text{minimize} \quad J(\mathbf{A}_q, \mathbf{A}_r) = \mathrm{tr}\left(\mathbf{A}_q^T \mathbf{M}_{qq} \mathbf{A}_q + \mathbf{A}_r^T \mathbf{M}_{qr} \mathbf{A}_r - 2\mathbf{A}_q^T \mathbf{M}_{qr} \mathbf{A}_r\right), \tag{13}$$

$$\text{s.t} \quad \mathbf{A}_q^T \mathbf{C}_q \mathbf{A}_q + \mathbf{A}_r^T \mathbf{C}_r \mathbf{A}_r = \mathbf{I}. \tag{14}$$

For Eq.(13) $\mathbf{M}_{qq} = \mathbf{X}_q \mathbf{R}_q \mathbf{X}_q^T$, $\mathbf{M}_{rr} = \mathbf{X}_r \mathbf{R}_r \mathbf{X}_r^T$, and $\mathbf{M}_{qr} = \mathbf{X}_q \mathbf{R}_q \mathbf{X}_r^T$. While for Eq.(14), $\mathbf{C}_q = \frac{1}{N_q} \mathbf{X}_q \mathbf{X}_q^T$ and $\mathbf{C}_r = \frac{1}{N_r} \mathbf{X}_r \mathbf{X}_r^T$ are the covariance matrices.

To solve the optimization problem, we introduce the matrices

$$\mathbf{M} = \begin{pmatrix} \mathbf{X}_q \mathbf{R}_q \mathbf{X}_q^T & -\mathbf{X}_q \mathbf{U} \mathbf{X}_r^T \\ -\mathbf{X}_r \mathbf{U}^T \mathbf{X}_q^T & \mathbf{X}_r \mathbf{R}_r \mathbf{X}_r^T \end{pmatrix} \qquad \mathbf{A} = \begin{pmatrix} \mathbf{A}_q \\ \mathbf{A}_r \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \mathbf{C}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_r \end{pmatrix} \tag{15}$$

According to Eq.(13), Eq.(14), and Eq.(15), the optimization problem can be written as

$$\mathbf{A} = \underset{\mathbf{A}^T \mathbf{C} \mathbf{A} = \mathbf{I}}{\operatorname{argmin}} \mathbf{A}^T \mathbf{M} \mathbf{A}, \tag{16}$$

where both $\mathbf{M}$ and $\mathbf{C}$ are $(d_q + d_r) \times (d_q + d_r)$ symmetric matrices.

[1] Here we assume that the samples $\mathbf{X}_q$ and $\mathbf{X}_r$ have zero mean vectors, otherwise, we can first shift them by subtracting the mean vectors.

To solve the constraint optimization problem, we have the following lemma

Lemma 1. *The matrix* $\mathbf{A}$ *satisfies* $\mathbf{ACA}^T = \mathbf{I}$ *where* $\mathbf{C}$ *is symmetric,* ***if and only if*** $\mathbf{A}$ *can be written as* $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}$ *where columns of* $\mathbf{V}$ *are eigenvectors and* $\mathbf{\Lambda}$ *are diagonal matrix of eigenvalues satisfying* $\mathbf{CV} = \mathbf{V\Lambda}$*, and* $\mathbf{U}$ *are orthogonal matrix satisfying* $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

The lemma suggests a two-stage diagonalization scheme to obtain the optimal solution. In the first stage, we solve the $\mathbf{V}$ and $\mathbf{\Lambda}$ by eigenvalue decomposition on $\mathbf{C}$ and compute the whitening transform $\mathbf{W} = \mathbf{V}\mathbf{\Lambda}^{-\frac{1}{2}}$. It can be easily shown that $\mathbf{T}^T\mathbf{CT} = \mathbf{I}$. Considering that $\mathbf{C}$ is a block-diagonal matrix, it be accomplished by eigen-decomposition on $\mathbf{C}_q$ and $\mathbf{C}_r$ respectively as $\mathbf{C}_q = \mathbf{V}_q\mathbf{\Lambda}_q\mathbf{V}_q^T$ and $\mathbf{C}_r = \mathbf{V}_r\mathbf{\Lambda}_r\mathbf{V}_r^T$. When the dimensions of $\mathcal{X}_q$ and $\mathcal{X}_r$ are high, the covariance matrices may become nearly singular and incur instability. To stabilize the solution, we approximate the covariance by discarding the eigenvalues near zero and the corresponding eigenvectors as follows:

$$\widetilde{\mathbf{C}}_q = \widetilde{\mathbf{V}}_q\widetilde{\mathbf{\Lambda}}_q\widetilde{\mathbf{V}}_q^T \qquad \widetilde{\mathbf{C}}_r = \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r\widetilde{\mathbf{V}}_r^T \tag{17}$$

Subsequently, $\mathbf{T}$ can be obtained by $\mathbf{T} = \begin{pmatrix} \widetilde{\mathbf{V}}_q\widetilde{\mathbf{\Lambda}}_q^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r^{-\frac{1}{2}} \end{pmatrix}$.

Then the learning objective is transformed to be

$$\mathbf{U} = \underset{\mathbf{U}}{\operatorname{argmin}}\, \mathbf{U}^T\left(\mathbf{T}^T\mathbf{MT}\right)\mathbf{U}, \qquad \text{s.t } \mathbf{U}^T\mathbf{U} = \mathbf{I}, \tag{18}$$

In the second stage we solve $\mathbf{U}$ by eigen-decomposition on the matrix $\mathbf{M}_W = \mathbf{T}^T\mathbf{MT}$ and taking the eigenvectors associated with the smallest eigenvalues of $\mathbf{M}_W$, then $\mathbf{A} = \mathbf{TU}$. Exploiting the fact that $\mathbf{T}$ is block-diagonal, we further simplify the computation by partitioned matrix multiplication. The whole procedure is summarized in Table 1.

Table 1. The Procedure of Solving the Linear Transform

1. Compute $\mathbf{R}_q$, $\mathbf{R}_r$ and $\mathbf{U}$ as in section 2.2.
2. Compute $\mathbf{M}_{qq} = \mathbf{X}_q\mathbf{R}_q\mathbf{X}_q^T$, $\mathbf{M}_{qr} = \mathbf{X}_q\mathbf{U}\mathbf{X}_r^T$ and $\mathbf{M}_{rr} = \mathbf{X}_r\mathbf{R}_r\mathbf{X}_r^T$.
3. Compute $\mathbf{C}_q = \frac{1}{N_q}\mathbf{X}_q\mathbf{X}_q^T$ and $\mathbf{C}_r = \frac{1}{N_r}\mathbf{X}_r\mathbf{X}_r^T$.
4. Solve $\widetilde{\mathbf{V}}_q$, $\widetilde{\mathbf{\Lambda}}_q$, $\widetilde{\mathbf{V}}_r$ and $\widetilde{\mathbf{\Lambda}}_r$ by performing eigenvalue-eigenvector analysis on $\mathbf{C}_q$ and $\mathbf{C}_r$ and removing the trailing eigenvalues and corresponding eigenvectors.
5. Compute $\mathbf{T}_q = \widetilde{\mathbf{V}}_q\widetilde{\mathbf{\Lambda}}^{-\frac{1}{2}}$ and $\mathbf{T}_r = \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}^{-\frac{1}{2}}$. Denote their numbers of columns by $\widetilde{d}_q$ and $\widetilde{d}_r$.
6. Compute $\mathbf{M}_W = \begin{pmatrix} \mathbf{T}_q^T\mathbf{M}_{qq}\mathbf{T}_q & -\mathbf{T}_q^T\mathbf{M}_{qr}\mathbf{T}_r \\ -\mathbf{T}_r^T\mathbf{M}_{qr}^T\mathbf{T}_q & \mathbf{T}_r^T\mathbf{M}_r r\mathbf{T}_r \end{pmatrix}$.
7. Solve $\mathbf{U}$ by taking the eigenvectors corresponding to the d least eigenvalues of $\mathbf{M}_W$.
8. Denote the first $\widetilde{d}_q$ rows of $\mathbf{U}$ by $\mathbf{U}_q$ and the rest $\widetilde{d}_r$ rows by $\mathbf{U}_r$. Then we have $\mathbf{A}_q = \mathbf{T}_q\mathbf{U}_q$ and $\mathbf{A}_r = \mathbf{T}_r\mathbf{U}_r$.

3 Kernelized Common Discriminant Feature Extraction

Kernel-based learning is often used to exploit the nonlinearity of the sample space. The core principle is to map the samples to a Hilbert space with much higher dimension or even infinite dimension so that the inner product structure of that space reflects the desirable similarity. Suppose the original sample space is denoted by $\mathcal{X}$ and a positive definite kernel is defined on it by $k : \mathcal{X} \times \mathcal{X} \to R$. For a set of observed samples: $\{\mathbf{x}_i\}_{i=1}^n$, the $n \times n$ *Gram matrix* is given by $\mathbf{K}$ with $\mathbf{K}(i,j) = k(\mathbf{x}_i, \mathbf{x}_j)$.

According to the kernel theory, each positive definite kernel k induces a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ satisfying that for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle = k(\mathbf{x}_1, \mathbf{x}_2)$. With this kernel trick, we can compute the inner product in the original space without explicitly evaluating the feature map.

Given the Hilbert space, we can extract the features by projecting the high-dimensional mapping to a lower-dimensional feature space. Assume the basis of the projection is a linear combinations of the Hilbert mappings of the training samples. Denote $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]$, then we have $\mathbf{P} = \mathbf{\Phi}\mathbf{A}$, where $\mathbf{A}$ is an $n \times d$ matrix storing the expansion coefficients and d is the dimension of the final feature space. Then for any sample $\mathbf{x} \in \mathcal{X}$, it is transformed to

$$\mathbf{y} = \mathbf{P}^T \phi(\mathbf{x}) = \mathbf{A}^T \mathbf{\Phi}^T \phi(\mathbf{x}) = \mathbf{A}^T \mathbf{k}(\mathbf{x}), \tag{19}$$

where $\mathbf{k}(\mathbf{x}) = [\phi(\mathbf{x}_1, \mathbf{x}), \phi(\mathbf{x}_1, \ldots, \phi(\mathbf{x}_n, \mathbf{x})]^T$. Specially, for the training set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, the matrix of the transformed vectors can be expressed as

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] = [\mathbf{P}\phi(\mathbf{x}_1), \mathbf{P}\phi(\mathbf{y}_2), \ldots, \mathbf{P}\phi(\mathbf{y}_n)] = \mathbf{A}^T \mathbf{K}. \tag{20}$$

Actually, the learning of Common Discriminant Feature Extraction relies on inner products, thus it can be extended to the nonlinear case by kernel theory. Denote the Gram matrices for the query samples and the reference samples by $\mathbf{K}_q$ and $\mathbf{K}_r$, and the coefficient expansion matrices for transform operators by $\mathbf{A}_q$ and $\mathbf{A}_r$. According to Eq.(20), we have the feature vectors for the training set expressed as follows:

$$\mathbf{Y}_q = \mathbf{A}_q^T \mathbf{K}_q \qquad \mathbf{Y}_r = \mathbf{A}_r^T \mathbf{K}_r. \tag{21}$$

Then from Eq.(10), the joint objective function can be written by

$$J(\mathbf{A}_q, \mathbf{A}_r) = \mathrm{tr}(\mathbf{A}_q^T \mathbf{K}_q \mathbf{R}_q \mathbf{K}_q^T \mathbf{A}_q + \mathbf{A}_r^T \mathbf{K}_r \mathbf{R}_r \mathbf{K}_r^T \mathbf{A}_r - 2\mathbf{A}_q^T \mathbf{K}_q \mathbf{U} \mathbf{K}_r^T \mathbf{A}_r) \tag{22}$$

$$\text{s.t} \quad \mathbf{A}_q \left(\frac{1}{N_q} \mathbf{K}_q \mathbf{K}_q^T \right) \mathbf{A}_q^T + \mathbf{A}_r^T \left(\frac{1}{N_r} \mathbf{K}_r, \mathbf{K}_r^T \right) \mathbf{A}_r = \mathbf{I}. \tag{23}$$

Comparing Eq.(13) and Eq.(23), we see that the mathematical form of the optimization problem is essentially the same, except that the matrices $\mathbf{X}_q$ and $\mathbf{X}_r$ are replaced by the Kernel Gram matrices $\mathbf{K}_q$ and $\mathbf{K}_r$. Thus the optimization procedure derived above is also applicable here.

4 Multi-mode Framework

In practical systems, the reference images are often captured in a controlled condition, while the query images on spot are subject to significant variation of illumination and pose. To address this problem, we develop a Multi-Mode Framework. For each query mode, we learn a common feature space for comparing the query samples in that mode and the reference samples. Here, we denote the transform matrices for the k-th mode by $\mathbf{A}_{qk}$ and $\mathbf{A}_{rk}$.

Considering that uncertainty may arise when we judge which mode a query sample belongs to, we adopt a soft fusion scheme. In the scheme, the *fused distance* is introduced to measure the dissimilarity between the query samples and the reference samples, which is a belief-based weighted combination of the distance values evaluated in the common spaces for different modes. We denote the belief that the i-th query sample belongs to the k-th mode by b_{ik}, and denote the features of the i-th query sample and the j-th reference sample in the common space for the k-th mode by $\mathbf{y}_{ik}^{(q)} = \mathbf{A}_{qk}^T \mathbf{x}_i^{(q)}$ and $\mathbf{y}_{jk}^{(r)} = \mathbf{A}_{rk}^T \mathbf{x}_i^{(r)}$ respectively, then the fused distance is given by

$$d(\mathbf{x}_i^{(q)}, \mathbf{x}_j^{(r)}) = \sum_{k=1}^{M} b_{ik} ||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2 \qquad \text{s.t} \sum_{k=1}^{M} b_{ik} = 1. \tag{24}$$

When the belief values for training samples are known, for a new query sample $\mathbf{x}$, its belief values w.r.t to the modes can be computed by smooth interpolation from the training samples adjacent to it. We re-formulate the learning objective with the following extensions:

1) Evaluate the empirical separability based on fused distance: $J_e = \sum_{i=1}^{N_q} \sum_{j=1}^{N_r} u_{ij} d(\mathbf{x}_i^{(q)}, \mathbf{x}_j^{(r)})$;
2) The local consistency comprises the local consistency of transforms for all modes;
3) Each query samples in the training set corresponds to M belief values. To ensure each mode covers a continuous and smooth region in the sample space so that the computation of beliefs for new samples is stable, we further enforce the local consistency on the belief values: $J_l^{(b)} = \sum_{i=1}^{N_q} \sum_{j=1}^{N_q} v_{ij}^{(q)} \sum_{k=1}^{M} (b_{ik} - b_{jk})^2$.
Consequently, the multimode formulation of the learning objective is derived as follows:

$$J = J_e + \beta \sum_{i=1}^{M} (J_l^{(q)} + J_l^{(r)}) + \gamma J_l^{(b)}, \tag{25}$$

where γ controls the contribution of the local consistency of beliefs. Eq.(25) can be expanded as follows:

$$\begin{aligned} J = & \sum_{i=1}^{N_q} \sum_{j=1}^{N_r} u_{ij} \sum_{k=1}^{M} b_{ik} ||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2 + \sum_{k=1}^{M} \sum_{i=1}^{N_q} \sum_{j=1}^{N_q} v_{ij}^{(q)} ||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(q)}||^2 \\ & + \sum_{k=1}^{M} \sum_{i=1}^{N_r} \sum_{j=1}^{N_r} v_{ij}^{(r)} ||\mathbf{y}_{ik}^{(r)} - \mathbf{y}_{jk}^{(r)}||^2 + \sum_{i=1}^{N_q} \sum_{j=1}^{N_q} v_{ij}^{(q)} \sum_{i=1}^{M} (b_{ik} - b_{jk})^2. \end{aligned} \tag{26}$$

Based on the generalized formulation, we derive the optimization scheme by alternate optimizing the transform matrices and the belief values.

1) Optimizing Transform Matrices. Denote $J_A = J_e + \beta \sum_{i=1}^{M}(J_l^{(q)} + J_l^{(r)})$, since $J_l^{(b)}$ does not relate to the features, with the belief values given, we can obtain the optimal transform matrices by minimizing J_A. Rearranging the order of sums, we can write it by

$$J_A = \sum_{k=1}^{M}\left\{\sum_{i=1}^{N_q}\sum_{j=1}^{N_r} b_{ik}u_{ij}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2 + \sum_{i=1}^{N_q}\sum_{j=1}^{N_q} v_{ij}^{(q)}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(q)}||^2 + \sum_{i=1}^{N_r}\sum_{j=1}^{N_r} v_{ij}^{(r)}||\mathbf{y}_{ik}^{(r)} - \mathbf{y}_{jk}^{(r)}||^2\right\}. \tag{27}$$

Thus J_A can be decomposed into

$$J_A = \sum_{k=1}^{M} J_k(\mathbf{A}_k^{(q)}, \mathbf{A}_k^{(r)}) \tag{28}$$

$$J_k(\mathbf{A}_k^{(q)}, \mathbf{A}_k^{(r)}) = \sum_{i=1}^{N_q}\sum_{j=1}^{N_r} b_{ik}u_{ij}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2 + \sum_{i=1}^{N_q}\sum_{j=1}^{N_q} v_{ij}^{(q)}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(q)}||^2 + \sum_{i=1}^{N_r}\sum_{j=1}^{N_r} v_{ij}^{(r)}||\mathbf{y}_{ik}^{(r)} - \mathbf{y}_{jk}^{(r)}||^2. \tag{29}$$

Compare Eq.(6) and Eq.(29), we see that they share the same mathematical form except that u_{ij} is replaced by $b_{ik}u_{ij}$. Because J_k is solely determined by the features of the k-th mode, we can optimize $\mathbf{A}_k^{(q)}$ and $\mathbf{A}_k^{(r)}$ for each mode individually by the aforementioned procedure with the belief values fixed.

2) Optimizing Belief Values. Denote $J_B = J_e + \gamma J_l^{(b)}$, which is the part of objective depending on the belief values. With the transform matrices given, we can optimize the beliefs by minimizing J_B:

$$J_B = \sum_{i=1}^{N_q}\sum_{j=1}^{N_r}\sum_{k=1}^{M} u_{ij}\sum_{k=1}^{M} b_{ik}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2 + \sum_{i=1}^{N_q}\sum_{j=1}^{N_q}\sum_{k=1}^{M} v_{ij}^{(q)}\sum_{i=1}^{M}(b_{ik} - b_{jk})^2. \tag{30}$$

For succinctness, we denote $e_{ik} = \sum_{j=1}^{N_r} u_{ij}||\mathbf{y}_{ik}^{(q)} - \mathbf{y}_{jk}^{(r)}||^2$, then it can be simplified to

$$J_B = \sum_{i=1}^{N_q}\sum_{k=1}^{M} e_{ik}b_{ik} + \sum_{i=1}^{N_q}\sum_{j=1}^{N_q}\sum_{k=1}^{M} v_{ij}^{(q)}\sum_{i=1}^{M}(b_{ik} - b_{jk})^2. \tag{31}$$

We introduce the following notations: $\mathbf{E}$ is an $M \times N_q$ matrix with $\mathbf{E}(i,k) = e_{ik}$, $\mathbf{B}$ is an $M \times N_q$ matrix with $\mathbf{B}(i,k) = b_{ik}$, then the optimization problem can be written in a matrix form as

$$\mathbf{B} = \underset{\mathbf{B}}{\operatorname{argmin}}\, J_B = \underset{\mathbf{B}}{\operatorname{argmin}}\, \operatorname{tr}(\mathbf{E}^T\mathbf{B} + 2\mathbf{B}(\mathbf{D}_q - \mathbf{V}_q)\mathbf{B}^T), \quad \text{s.t } \mathbf{B}^T\mathbf{1}_M = \mathbf{1}_{N_q}. \tag{32}$$

Here $\mathbf{D}_q - \mathbf{V}_q$ is positive-semidefinite. This is a convex quadratic optimization program with linear constraint and can be efficiently solved by quadratic programming.

3) The whole procedure of optimization. We adopt the alternate optimization strategy in our framework. First we cluster all the query samples in the training set by Gaussian Mixture Model and set the initial belief values to be the posteriori evaluated by GMM. After that, we optimize the transform matrices for each mode based on Eq.(29) and the belief values based on Eq.(32) alternately until convergence.

5 Experiments

Experiment Settings

We conduct experiments in two inter-modality recognition applications.

1) Infrared-optical recognition. The reference images are captured by optical cameras with controlled illumination condition, while the query images are acquired in an uncontrolled environment. To cope with the adverse illumination condition, we use infrared cameras to capture the query images. In our experiment, two configurations are constructed to test our algorithms. Both configurations share the same set of reference samples. The reference set consists of 64 samples from 16 persons with each person having 4 samples. In the first configuration, we select 800 images with mild expression variation to form the query set. The second configuration is a much more challenging one, which consists of 1600 images subject to significant pose and illumination variation. Some examples of the images are displayed in fig.1. It can be seen that the infrared images are seriously blurred and distorted due to the limitation of infrared imaging.

2) Sketch-photo recognition. The reference set is composed of 350 images from FERET face database[16]. The 350 images represent 350 different persons. The query set comprises 700 sketches composed by artists. Each person has 2 samples in the query set. Fig.2 shows some examples of the photos and the corresponding sketches. We can see that the sketches present greatly different characteristics from the photos. In addition, some texture information is lost in the sketches.

All the photos are normalized to reduce the influence of interference factors. For each image, we first perform affine transformation on it to fix the eye centers and mouth center of the face to standard positions. Then we crop it to the size of 64×72. After that we apply histogram equalization and mask the background region using a face-shape mask. After preprocessing, we obtain the original vector representation for each image by scanning the 4114 remaining pixels to a vector. To accelerate the process of training and testing and suppress the noise, we employ PCA to reduce the space dimension and preserve 0.98% of the energy in the principal space.

Experiment Results

1) We first investigate how the selection of parameters α and β affects the generalization performance. In the experiments, we find that the performance is not sensitive to the α when α ranges from 0.2 to 2. However, the parameter β significantly influence the results. Fig.5, fig.7 and fig.9 show the change of performance w.r.t the number of features when β takes different values. We can see that when

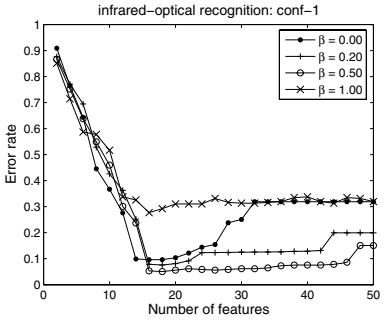

Fig. 5. Performance of CDFE in conf-1 of infrared-optical recognition

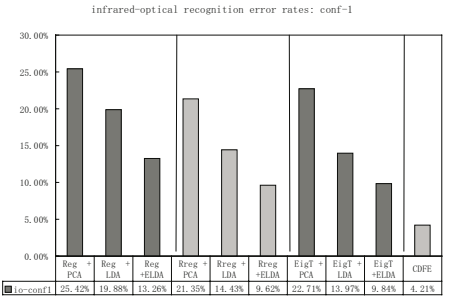

Fig. 6. Comparison of algorithms in conf-1 of infrared-optical recognition

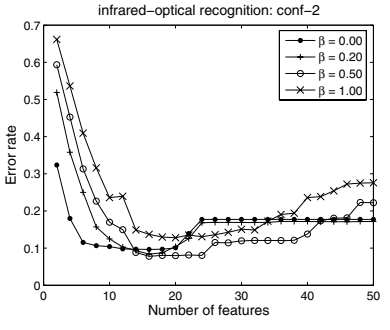

Fig. 7. Performance of CDFE in conf-2 of infrared-optical recognition

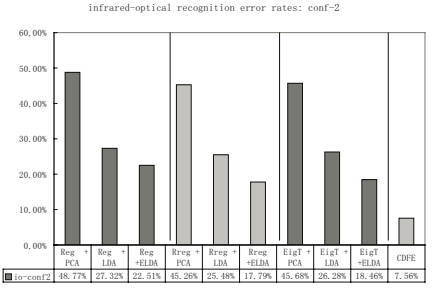

Fig. 8. Comparison of algorithms in conf-2 of infrared-optical recognition

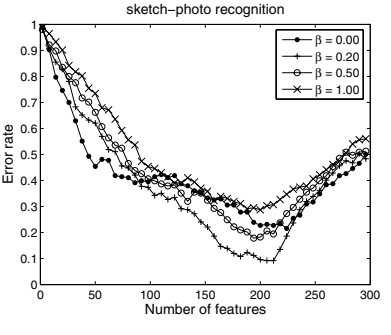

Fig. 9. Performance of CDFE in sketch-photo recognition

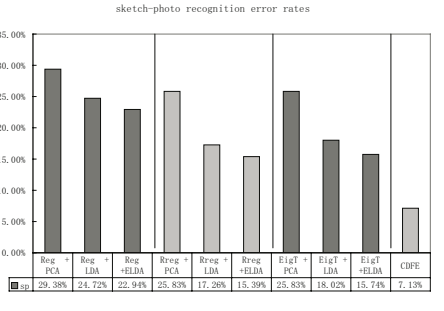

Fig. 10. Comparison of algorithms in sketch-photo recognition

$\beta = 0$, that is, the local consistency does not contribute to the formulation, the performance degrades drastically as the number of features increases. When β becomes larger, the change of performance becomes stable. However, if β is too large, the performance may degenerate. This is mainly due to over-smoothing. From the results, we can see that for infrared-optical recognition, the algorithm

	Reg. + KPCA	Reg. + KLDA	RReg. + KPCA	RReg. + KLDA	EigT. + KPCA	EigT. + KLDA	CDFE
infrared-optical. conf-1	24.3%	15.6%	20.4%	8.34%	22.7%	9.56%	1.98%
infrared-optical. conf-2	50.6%	21.8%	47.9%	15.0%	45.7%	15.3%	4.42%
sketch-optical.	32.2%	20.8%	24.3%	11.7%	25.8%	12.8%	5.43%

Fig. 11. Comparison of the algorithms with kernelized features

achieves best performance when $\beta = 0.5$, while for sketch-photo recognition, the algorithm achieves best performance when $\beta = 1.0$. The analysis above indicates the important role of local consistency for the generalization ability.

2) We compare the common discriminant feature extraction (CDFE) with other approaches for inter-modality recognition. In previous works, it is typical to first convert the query images to the reference modality and then apply conventional algorithms to classify the converted sample. In the experiments, we test the combination of three conversion methods (linear regression (Reg), ridge regression (RReg), and Eigentransformation (EigT)[9]) and three feature extraction methods (PCA[17], LDA[1], and Enhanced LDA[4]). The results are illustrated in fig.6, fig.8, and fig.10 for the infrared-optical recognition and the sketch-photo recognition respectively. It can be seen from the results that the CDFE consistently outperforms the other methods by a large margin. In all the configurations, CDFE at least reduces the error rate by half compared with the most competitive methods in conventional approaches.

3) We test the kernelized extension of the CDFE and compare it with the conversion-classification paradigm. For fair comparison, in the traditional approach, we also use kernelized method to extract features. Here, we test Kernel PCA and Kernel LDA. Gaussian kernel is used in the testing. The results are listed in fig.11. All the results given in the table are the best performances obtained through cross-validation. We can see our algorithm outperforms the conventional ones by a surprisingly large margin. In our view, such a remarkable improvement is owing to incorporation of local consistency, which on one hand fully exploits the potency of kernel method, on the other hand effectively controls the complexity of the operator.

4) We finally test the multi-mode framework in the conf-2 of infrared-optical recognition. In this configuration, due to diverse poses and illumination conditions, there are multiple modes in the sample distribution. In our experiments, the error rate decreases when we increase the number of modes. The lowest error rate 3.25% is attained when $M = 5$. Compared to the single mode case, in which error rate is 7.56%, it is an encouraging improvement.

6 Conclusion

In this paper, we studied the inter-modality face recognition problem. We proposed a new notion of common discriminant feature space and formulated the learning objective with local consistency. In the extensive experiments, our algorithms have achieved significant improvement over conventional methods.

Acknowledgement

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region and a joint grant (N_CUHK409-03) from HKSAR RGC and China NSF. The work was done in The Chinese University of Hong Kong.

References

1. P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on PAMI **19**(7) (1997) 711–720
2. W. Zhao, R. Chellappa, A. Krishnaswamy: Discriminant Analysis of Principal Components for Face Recognition. Proc. of FGR'98 (1998)
3. J. Yang, F. Frangi, J. Yang, D. Zhang, Z. Jin: KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition. IEEE Trans. on PAMI **27**(2) (2005) 230–244
4. C. Liu, H. Wechsler: Enhanced Fisher Linear Discriminant Models for Face Recognition. Proc. of CVPR'98 (1998)
5. X. Wang, X. Tang: A Unified Framework for Subspace Face Recognition. IEEE Trans. on PAMI **26**(9) (2004) 1222–1228
6. X. Wang, X. Tang: Unified Subspace Analysis for Face Recognition. In: Proc. of ICCV'03. (2003)
7. X. Wang, X. Tang: Dual-Space Linear Discriminant Analysis for Face Recognition. In: Proc. of CVPR'04. (2004)
8. X. Tang, X. Wang: Face Sketch Synthesis and Recognition. In: Proc. of ICCV'03. (2003)
9. X. Tang, X. Wang: Face Sketch Recognition. IEEE Trans. CSVT **14**(1) (2004) 50–57
10. V.N. Vapnik: Statistical Learning Theory. John Wiley and Sons, Inc. (1998)
11. D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf: Learning with Local and Global Consistency. In: Proc. of NIPS'04. (2004)
12. X. He, S. Yan, Y. Hu, H. Zhang: Learning a Locality Preserving Subspace for Visual Recognition. In: Proc. of ICCV'03. (2003)
13. X. He, Y. Hu, P. Niyogi, H. Zhang: Face Recognition Using Laplacianfaces. IEEE Trans. PAMI **27**(3) (2005) 328–340
14. M. Belkin, P. Niyogi: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: Proc. of NIPS'01. (2001)
15. T. Kim, J. Kittler: Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image. IEEE Trans. on PAMI **27**(3) (2005) 318–327
16. P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss: The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Trans. PAMI (2000) 1090–1104
17. M. Turk, A. Pentland: Eigenfaces for Recognition. J. Cogn. Neuro. **3**(1) (1991) 71–86