JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 29, 907-924 (2013)

DOI:10.1688/JISE.2013.29.5.7

Facial Expression Recognition using Spectral Supervised Canonical Correlation Analysis*

SONG GUO, QIUQI RUAN, ZHAN WANG AND SHUAI LIU
Institute of Information Science
Beijing Jiaotong University
Beijing, 100044 P.R. China

Feature extraction plays an important role in facial expression recognition. Canonical correlation analysis (CCA), which studies the correlation between two random vectors, is a major linear feature extraction method based on feature fusion. Recent studies have shown that facial expression images often reside on a latent nonlinear manifold. However, either CCA or its kernel version KCCA, which is globally linear or nonlinear, cannot effectively utilize the local structure information to discover the low-dimensional manifold embedded in the original data. Inspired by the successful application of spectral graph theory in classification, we proposed spectral supervised canonical correlation analysis (SSCCA) to overcome the shortcomings of CCA and KCCA. In SSCCA, we construct an affinity matrix, which incorporates both the class information and local structure information of the data points, as the supervised matrix. The spectral feature of covariance matrices is used to extract a new combined feature with more discriminative information, and it can reveal the nonlinear manifold structure of the data. Furthermore, we proposed a unified framework for CCA to offer an effective methodology for non-empirical structural comparison of different forms of CCA as well as providing a way to extend the CCA algorithm. The correlation feature extraction power is then proposed to evaluate the effectiveness of our method. Experimental results on two facial expression databases validate the effectiveness of our method.

Keywords: spectral supervised canonical correlation analysis, spectral classification, feature fusion, feature extraction, facial expression recognition

1. INTRODUCTION

Facial expression conveys visual human emotions, which makes the facial expression recognition (FER) plays an important role in human-computer interaction, image retrieval, synthetic face animation, video conferencing, human emotion analysis [1, 2]. Due to its wide range of applications, FER has attracted much attention in recent years. Generally speaking, a FER system consists of three major components: face detection, facial expression feature extraction and facial expression classification [1, 2]. Since appropriate facial expression representation can effectively alleviate the complexity of the design of classification and improve the performance of the FER system, most researches currently concentrate on how to extract effective facial expression features.

A variety of methods have been proposed for facial expression feature extraction [3-7], and there are generally two common approaches: single feature extraction and fea-

Received July 6, 2011; revised November 11, 2011; accepted December 1, 2011.
Communicated by Tyng-Luh Liu.
* This paper was supported by the National Natural Science Foundation of China (Grant No. 60973060), Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 200800040008), Beijing Program (Grant No. YB20081000401) and the Fundamental Research Funds for the Central Universities (Grant No. 2011JBM022).

ture fusion. Single feature extraction is based on a particular method, *i.e.* principal component analysis (PCA) [3], fisher's linear discriminant (FLD) [4], locality preserving projection (LPP) [5], *etc.*, to obtain the facial expression feature. Feature fusion is based on the selected single features and a fusion strategy, *i.e.* serial strategy [8], parallel strategy [8], canonical correlation analysis (CCA) [7, 9], *etc.*, to create a new feature vector which is more effective for classification.

CCA is a major method of feature fusion [9], which seeks to utilize paired dataset to find projections from each feature space that maximize the correlation between the projected representations. However, as a linear feature fusion technique, CCA can only reveal the linear correlation relationship between two sets of data, and it fails to discover the nonlinear correlation relationship between them. In contrast, its kernel-based nonlinear version, KCCA [7, 10], can overcome such a drawback, and has been adopted in facial expression recognition [7].

In KCCA [7, 10], the original facial expression features sets are mapped into a higher, possibly infinite dimensional feature space via implicit nonlinear mapping, *i.e.* Φ: $x \mapsto \Phi(x)$ and Ψ: $y \mapsto \Psi(y)$, for X and Y respectively, then the traditional CCA is performed in the feature space using kernel trick [7, 10-12]. Therefore, a nonlinear problem in the original space can be transformed into another more possibly linear one in the feature space in order to discover the nonlinear correlation between the original data sets. However, the global kernelization of CCA cannot necessarily guarantee the transformation from a nonlinear problem into a linear one in the feature space, especially when the nonlinearity only exists in certain local spaces. Furthermore, the choice of kernel function and its parameter(s) according to different research problems is still a tough problem [11-13].

Recent studies show that facial expression images often reside on a low dimensional manifold [14, 16]. The analysis and recognition of different facial expressions will be facilitated on the manifold. Both CCA and KCCA can discover the low dimensional manifold where the high dimensional data lies, to some extent, but in a global way. They cannot effectively guarantee that the similar expressions locate on local neighborhoods on the manifold.

Our work here is motivated by the successful application of spectral graph theory in dimensionality reduction, *i.e.* locally linear embedding (LLE) [15, 16], LPP [5], spectral feature analysis (SFA) [17], *etc.* LLE [15, 16] which is proposed by Roweis and Saul, is based on the assumption that one point can be represented by the linear combination of its local neighbors and this linear relationship still holds in low dimensional manifold embedded in the ambient space. He and Niyogi also propose LPP [5] to optimally preserve the neighborhood structure of the data set based on spectral graph theory. Recently, SFA [17] is proposed to utilize the spectral feature of the affinity matrix to extract discriminative information for classification. These approaches mentioned above are nonlinear in nature, but they can be achieved using linear mapping technology, based on the idea that the global nonlinear structure is locally linear. Meanwhile, these approaches share a common characteristic that they utilize the specific local structure information to reveal the low dimensional manifold embedded in the original high dimensional space.

As in classification, the class labels of the data points are available. Utilizing the class information, a new supervised feature extraction method, named supervised canonical correlation analysis (SCCA), is proposed. SCCA can maximize the ratio of the

between-set covariance and the within-set covariance, which ensures that the fusion feature contains more discriminative information for classification.

Furthermore, inspired by the successful application of spectral graph theory in classification, we propose a more effective feature fusion method called spectral supervised canonical correlation analysis (SSCCA) to tackle the previously mentioned problems in CCA and KCCA. In SSCCA, we construct an affinity matrix, which incorporates both the class information and local structure information of the data points, as the supervised matrix. The spectral feature of covariance matrices (within-set covariance matrices and between-set covariance) is used to extract a new combined feature, which can be more discriminative, and the local structure information existed in the original data points can be preserved in the feature space.

Moreover, we provide a unified framework for CCA to offer an effective methodology for non-empirical structural comparison of different forms of CCA as well as providing a way to extend the CCA algorithm. The correlation feature extraction power is also introduced to evaluate the performance of different forms of CCA.

Our method is particularly suitable for classification because of the following properties:

(1) SSCCA utilizes the class information of the data points to construct the affinity matrix. Class information is important for classification, so the features extracted from SSCCA have more powerful discriminative information for classification.
(2) SSCCA can reveal the intrinsic nonlinear manifold structure hidden in the original data. By incorporating the linear structure information of local neighbor, the global nonlinearity structure can be fully displayed in the low-dimensional manifold.
(3) SSCCA can have more correlation feature extraction power. The correlation feature extraction power is directly related to the discriminating power in recognition problem, so the SSCCA algorithm is more suitable for classification.
(4) SSCCA can effectively extract the fusion features from the testing samples as well as the training samples.

The rest of the paper is organized as follows. Section 2 introduces some related works. Section 3 describes SCCA and SSCCA in detail. Section 4 provides a unified framework of CCA, and the correlation feature extraction power is also introduced in this section. Experimental results are presented in section 5. Finally, the conclusions are drawn in section 6.

2. RELATED WORKS

2.1 Canonical Correlation Analysis (CCA)

CCA is a multivariate statistical analysis method that studies the correlation problem of two multidimensional random variables. It converts the correlation research of two multidimensional random variables into that of a few pairs of unrelated variables [18].

Concretely, given n pairs of centered data, (x_i, y_i), $x_i \in R^p$, $y_i \in R^q$, $i = 1, \ldots, n$ which

come from two information channels X and Y, respectively, CCA aims to find a pair of directions ω_x and ω_y so that the correlation between the projections $\omega_x^T x$ and $\omega_y^T y$ is maximized [18]. The correlation can be expressed as

$$\rho(x, y, \omega_x, \omega_y) = \frac{\operatorname{cov}(\omega_x^T x, \omega_y^T y)}{\sqrt{\operatorname{var}(\omega_x^T x)\operatorname{var}(\omega_y^T y)}} = \frac{\omega_x^T C_{xy} \omega_y}{\sqrt{\omega_x^T C_{xx} \omega_x} \bullet \sqrt{\omega_y^T C_{yy} \omega_y}} \tag{1}$$

where $C_{xx} = E(xx^T) = XX^T$ and $C_{yy} = E(yy^T) = YY^T$ are the within-set covariance matrices of X and Y respectively, $C_{xy} = E(xy^T) = XY^T$ is the between-set covariance matrix of X and Y.

Due to the scale invariance of ω_x and ω_y, the pair of projection (ω_x, ω_y) can be obtained by solving the following optimization problem

$$\begin{aligned} &\max_{\omega_x, \omega_y} \omega_x^T C_{xy} \omega_y \\ &s.t.\ \omega_x^T C_{xx} \omega_x = 1, \omega_y^T C_{yy} \omega_y = 1. \end{aligned} \tag{2}$$

Adopting the optimization strategy described in [9], this optimization problem can be solved by the following generalized eigenvalue problem

$$\begin{bmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \end{bmatrix} \tag{3}$$

where the eigenvalue λ is just the canonical correlation coefficient. Once the canonical correlation vector pairs (ω_x^i, ω_y^i), $i = 1, \ldots, d$, $d \le \min(p, q)$ are obtained, the following two linear transformations (4) and (5) can be adopted as the feature fusion strategy [9] to extract the fusion features for classification.

$$Z_1 = \begin{pmatrix} \omega_x^T x \\ \omega_y^T y \end{pmatrix} = \begin{pmatrix} \omega_x & 0 \\ 0 & \omega_y \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} \tag{4}$$

$$Z_2 = \omega_x^T x + \omega_y^T y = \begin{pmatrix} \omega_x \\ \omega_y \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix} \tag{5}$$

2.2 Spectral Classification

The spectral classification algorithm derives from spectral clustering [17]. Therefore, we first give a brief review on spectral clustering. Spectral clustering [19, 20] refers to a class of techniques which rely on the eigenstructure of affinity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity. Normalized cut (Ncut) algorithm [19] is one of the most effective spectral clustering methods.

Given the data set $\{x_1, x_2, \ldots, x_M\}$, with each $x_i \in R^d$, we can construct the affinity

matrix K by one of the ways described in [20], with its element k_{ij} represents the similarity between X_i and X_j. Considering the general c clustering problem, we separate the data set into c clusters $\{A_1, A_2, \ldots, A_c\}$, such that $A_i \cap A_j = \varnothing$, $i \neq j$ and $\cup_{i=1}^{c} A_i = A$. The Ncut criterion is defined as

$$Ncut(A_1, A_2, ..., A_c) = \sum_{i=1}^{c} \frac{cut(A_i, \overline{A_i})}{Vol(A_i)} \tag{6}$$

where $cut(A_i, \overline{A}_i) = \sum_{i \in A_i, j \in \overline{A}_i} k_{ij}$, $Vol(A_i) = \sum_{i \in A_i, j \in A} k_{ij}$.

Define Z as the indicator vector of data points for clustering, and then Eq. (6) can be relaxed to the following optimization problem:

$$J_Z = \frac{Z^T K Z}{Z^T D Z} \tag{7}$$

where D is a diagonal matrix with entries are row sums of the affinity matrix K.

The optimization problem of Eq. (7) can be solved using the Rayleigh-Ritz theorem, so the optimal Z is the solution of the following generalized eigenvalue problem: $KZ = \lambda DZ$. The first d columns of Z are kept for clustering. Each row of Z, $Z \in R^{n \times d}$, is the embedding projection of the corresponding original data pattern to spectral space, then the classical clustering algorithm such as k-means algorithm can be performed on the spectral feature space to cluster the points $\{z_1, z_2, \ldots, z_n\}$ into c clusters.

Besides for clustering, the spectral feature can also be used for classification, leading to the spectral classification [17]. However, traditional spectral clustering algorithms can only extract the spectral features of training set, and they cannot handle the testing data. To solve this problem, Kamvar and Klein [21] construct a Markov matrix to describe the transition probabilities between different data using the data similarities or supervisory information when the class labels are available. Eigen decomposition is then performed on the matrices constructed from both training and test sets. Wang and Zhang [17] indicate that spectral feature extraction is a special case of weighted kernel principal component analysis (WKPCA). Therefore, the spectral feature can be obtained from both training and testing sets using the kernel function.

Compared with the spectral clustering, the characteristics of the spectral classification lie on: (a) the class label information is available; (b) the spectral feature can be exacted from both training and testing sets; (c) a classifier is followed instead of a clustering algorithm.

3. SPECTRAL SUPERVISED CCA

In this section, we firstly propose SCCA by utilizing the class information. Secondly, the proposed feature fusion method SSCCA is demonstrated in detail.

3.1 Supervised CCA (SCCA)

As in classification, the class label is available, and the features which incorporate

the class information can be more discriminative for classification. By incorporating the class information to construct the supervised matrix, we propose a novel algorithm, called Supervised CCA (SCCA). The objective function of SCCA can be formulated as the following optimization problem

$$\begin{aligned} &\max_{\omega_x,\omega_y} \omega_x^T \widetilde{C_{xy}} \omega_y \\ &s.t.\ \omega_x^T \widetilde{C_{xx}} \omega_x = 1, \omega_y^T \widetilde{C_{yy}} \omega_y = 1. \end{aligned} \tag{8}$$

Incorporating the class information, $\widetilde{C_{xx}}$ and $\widetilde{C_{yy}}$ denote the within-set covariance matrix of X and Y respectively, $\widetilde{C_{xy}}$ denotes the between-set covariance matrix of X and Y, The details are described as follows.

Given two sets of mean-normalized samples X and Y, let

$$X = \left[x_1^1, ..., x_{n_1}^1, ..., x_1^c, ..., x_{n_c}^c \right] \in R^{p \times n}, \tag{9}$$

$$Y = \left[y_1^1, ..., y_{n_1}^1, ..., y_1^c, ..., y_{n_c}^c \right] \in R^{q \times n}, \tag{10}$$

$$e_{n_i} = [\underbrace{0, ...0}_{\sum_{j=1}^{i-1} n_j}, \underbrace{1, ...1}_{n_i}, \underbrace{0, ...0}_{\sum_{j=i+1}^{c} n_j}]^T \in R^n, \tag{11}$$

where x_j^i, y_j^i denote the jth sample in the ith class from X and Y respectively, n_i is the number of samples in the ith class of X (or Y) set, $\sum_{i=1}^{c} n_i = n$ and c is number of classes. e_{n_i} denotes that only the n_i samples from the ith class are 1, the remaining samples, *i.e.* the $\sum_{j=1}^{i-1} n_j$ samples from the j = 1st, …, (i − 1)th class and the $\sum_{j=i+1}^{c} n_j$ samples from the j = (i + 1)th, …, cth class, are 0. The within-set covariance with class information $\widetilde{C_{xx}}$ can be defined as

$$\widetilde{C_{xx}} = \sum_{i=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} x_k^i (x_l^i)^T = \sum_{i=1}^{c} (Xe_{n_i})(Xe_{n_i})^T = XSX^T, \tag{12}$$

where S denotes the supervised matrix, it can be defined as

$$S = \begin{pmatrix} 1_{n_1 \times n_1} & & & & \\ & \ddots & & & \\ & & 1_{n_i \times n_i} & & \\ & & & \ddots & \\ & & & & 1_{n_c \times n_c} \end{pmatrix} \in R^{n \times n}, \tag{13}$$

$$1_{n_i \times n_i} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \in R^{n_i \times n_i}. \tag{14}$$

As it reveals, S is block-diagonal to make sure that the samples from the same class stay close and the samples from different classes are away from each other. Furthermore, the elements lie on the block-diagonal of S are 1, the others are 0. Similarly, $\widetilde{C_{yy}} = YSY^T$, $\widetilde{C_{xy}} = XSY^T$.

The optimization problem of (8) is similar with that of (2), so it can be solved by adopting the same optimization strategy described in [9]. Once the basis vector pairs (ω_x^i, ω_y^i), $i = 1, \ldots, d$, $d \le \min(p, q)$ are obtained, the dimensionality reduction can be performed in the form of $\omega_x^T x$ and ω_y^T, and then the fusion strategy (4) or (5) can be adopted to extract the fusion features for classification.

3.2 Spectral Supervised CCA (SSCCA)

The supervised matrix constructed in SCCA can only explain the local structure of original data roughly, and it fails to accurately reveal the local neighborhood information of the samples in the same class. Fortunately, the goal of constructing affinity matrix in spectral classification is to model the specific of the local neighborhood relationship between the data points. Therefore, inspired by the successful application of spectral graph theory in classification, we propose a more effective feature fusion method called spectral supervised canonical correlation analysis (SSCCA). In the following two sections, the definition and derivation of SSCCA are described in detail respectively.

3.2.1 Definition of SSCCA

In SSCCA, we construct an affinity matrix, which incorporates both the class information and local structure information of the data points, as the supervised matrix. Utilizing the class information, the construction of the affinity matrix $K_{xx} = \left\{k_{ij}^x\right\}_{i,j=1}^n$ and $K_{yy} = \left\{k_{ij}^y\right\}_{i,j=1}^n$ can be defined as

$$k_{ij}^x = \begin{cases} \exp(-\left\|x_i - x_j\right\|^2 / t_x), \text{ if } x_i \text{ and } x_j \text{ are in the same class} \\ 0, \qquad\qquad \text{otherwise} \end{cases}, \tag{15}$$

$$k_{ij}^y = \begin{cases} \exp(-\left\|y_i - y_j\right\|^2 / t_y), \text{ if } y_i \text{ and } y_j \text{ are in the same class} \\ 0, \qquad\qquad \text{otherwise} \end{cases}. \tag{16}$$

Compared with SCCA, we replace S with the affinity matrix K_{xx} and K_{yy} as the supervised matrix in SSCCA. Accordingly, we have the following definition

$$\widetilde{C_{xx}} = XD_{xx}X^T, \widetilde{C_{yy}} = YD_{yy}Y^T, \ \widetilde{C_{xy}} = XK_{xy}Y^T \tag{17}$$

where $D_{xx}(D_{yy})$ is a diagonal matrix whose entries are row (or column, since K_{xx} or K_{yy} is symmetric) sum of the matrix $K_{xx}(K_{yy})$. $K_{xy} = K_{xx} + K_{yy}$ represents the supervised matrix of between-set correlation. By incorporating the class information and local neighbor information of samples, $\widetilde{C_{xx}}$ and $\widetilde{C_{yy}}$ denote the supervised within-set covariance matrices of X and Y respectively, $\widetilde{C_{xy}}$ denotes the supervised between-set covariance matrix of

X and *Y*.

The canonical correlation coefficient of SSCCA can be expressed as

$$\underset{\text{SSCCA}}{\rho}(x,y,\omega_x,\omega_y)=\frac{\omega_x^T XK_{xy}Y^T\omega_y}{\sqrt{\omega_x^T XD_{xx}X^T\omega_x\omega_y^T YD_{yy}Y^T\omega_y}}=\frac{\omega_x^T\widetilde{C_{xy}}\omega_y}{\sqrt{\omega_x^T\widetilde{C_{xx}}\omega_x\omega_y^T\widetilde{C_{yy}}\omega_y}}. \tag{18}$$

SSCCA aims to seek pairs of projection (ω_x, ω_y) such that the canonical correlation coefficient $\underset{\text{SSCCA}}{\rho}(x,y,\omega_x,\omega_y)$, which is the ratio of the between-set covariance and the within-set covariance, is maximized. Therefore, the objective function of SSCCA is the maximization of $\underset{\text{SSCCA}}{\rho}(x,y,\omega_x,\omega_y)$, it can be expressed as

$$J(\omega_x,\omega_y)=\arg\max_{\omega_x,\omega_y}\frac{\omega_x^T XK_{xy}Y^T\omega_y}{\sqrt{\omega_x^T XD_{xx}X^T\omega_x\omega_y^T YD_{yy}Y^T\omega_y}}. \tag{19}$$

From the optimization problem described in Eq. (19), we can see that it has the similar form with the optimization problem of spectral classification stated in Eq. (7). Therefore, it can be seen as a dualistic expansion of Eq. (7). The solving process of SSCCA will be detailed in the following section.

3.2.2 Derivation of SSCCA

With the introduction of $\widetilde{C_{xx}}, \widetilde{C_{yy}}$ and $\widetilde{C_{xy}}$, the maximization of $\underset{\text{SSCCA}}{\rho}(x,y,\omega_x,\omega_y)$ in Eq. (19) can be reformulated as the optimization problem in Eq. (8). Then it can be further solved by the following generalized eigenvalue problem:

$$\begin{bmatrix} 0 & \widetilde{C_{xy}} \\ \widetilde{C_{yx}} & 0 \end{bmatrix}\begin{bmatrix}\omega_x\\ \omega_y\end{bmatrix}=\lambda\begin{bmatrix}\widetilde{C_{xx}} & 0\\ 0 & \widetilde{C_{yy}}\end{bmatrix}\begin{bmatrix}\omega_x\\ \omega_y\end{bmatrix}. \tag{20}$$

Since $\widetilde{C_{xx}}$ and $\widetilde{C_{yy}}$ in Eq. (17) are both positive definite, we can obtain

$$\widetilde{C_{xy}}\widetilde{C_{yy}}^{-1}\widetilde{C_{yx}}\omega_x=\lambda^2\widetilde{C_{xx}}\omega_x, \tag{21}$$

$$\widetilde{C_{yx}}\widetilde{C_{xx}}^{-1}\widetilde{C_{xy}}\omega_y=\lambda^2\widetilde{C_{yy}}\omega_y. \tag{22}$$

Given $\omega_x=\widetilde{C_{xx}}^{-1/2}u, \omega_y=\widetilde{C_{yy}}^{-1/2}v,$ multiplying the left sides of Eqs. (21) and (22) by $\widetilde{C_{xx}}^{-1/2}$ and $\widetilde{C_{yy}}^{-1/2}$ respectively, we obtain

$$\widetilde{C_{xx}}^{-1/2}\widetilde{C_{xy}}\widetilde{C_{yy}}^{-1}\widetilde{C_{yx}}\widetilde{C_{xx}}^{-1/2}u=\lambda^2u, \tag{23}$$

$$\widetilde{C}_{yy}^{-1/2}\widetilde{C}_{yx}\widetilde{C}_{xx}^{-1}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1/2}v = \lambda^2 v. \tag{24}$$

For simplifying the further deduction, we define

$$\begin{aligned} G_x &= \widetilde{C}_{xx}^{-1/2}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1}\widetilde{C}_{yx}\widetilde{C}_{xx}^{-1/2}, \\ G_y &= \widetilde{C}_{yy}^{-1/2}\widetilde{C}_{yx}\widetilde{C}_{xx}^{-1}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1/2}. \end{aligned} \tag{25}$$

Let $H = \widetilde{C}_{xx}^{-1/2}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1/2}$, then $G_x = HH^T$, $G_y = H^TH$. We can find that G_x and G_y have the same nonzero eigenvalues $\lambda_1^2, \lambda_2^2, ..., \lambda_r^2, r = rank(\widetilde{C}_{xy})$. Applying singular value decomposition (SVD) theorem to matrix H, we get $H = \sum_{i=1}^{r} \lambda_i u_i v_i^T$, where $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_r^2$ are the nonzero eigenvalues of G_x and G_y, u_i and $v_i (i = 1, 2, \ldots, r)$ are the normalized orthogonal eigenvectors of G_x and G_y corresponding to the nonzero eigenvalue λ_i^2 respectively. It would be important to note that the derivation presented here is motivated by [9].

Furthermore, the spectral of H, *i.e.* $\lambda_1, \lambda_2, ..., \lambda_r, r = rank(\widetilde{C}_{xy})$ are just the canonical correlation coefficients of SSCCA. We can obtain the basis vector pairs (ω_x^i and ω_y^i) from the spectral feature vectors u_i and $v_i (i = 1, 2, \ldots, r)$ by $\omega_x^i = \widetilde{C}_{xx}^{-1/2}u_i, \omega_y^i = \widetilde{C}_{yy}^{-1/2}v_i$, $i = 1, 2, \ldots, r$, and then either the fusion strategy (4) or (5) can be adopted to extract the fusion feature in the form of $\omega_x^T x$ and ω_y^T. By incorporating the class information and local structure information, the combination feature extracted from our algorithm can be more discriminative and the local structure information of the original data space can be preserved in the feature space. We will validate the effectiveness of SSCCA in the FER experiments.

4. ANALYSIS AND DISCUSSION

In this section, we first present a unified framework for CCA. Then the correlation feature extraction power is introduced to evaluate the performance of different forms of CCA.

4.1 A Unified Framework for CCA

The CCA algorithm and all of its derivative algorithms can be expressed in a unified framework. All the derivative algorithms of CCA, including CCA algorithm itself, follow the same procedure, which can be explained as follows: extract two groups of feature vectors of the same problem; establish the correlation criterion function under certain conditions; extract uncorrelated canonical variates according to this criterion. Therefore, the correlation of two sets of variables can be studied through a few pairs of canonical variates.

Given two sets of mean-normalized multivariable vectors $X \in R^{p \times n}$, $Y \in R^{q \times n}$, pairs of basis vector (ω_x, ω_y) are the projection directions of X and Y respectively. The objective function of the unified framework of CCA is to maximize the correlation between the projections $\omega_x^T x$ and $\omega_y^T y$, so it can be described by Eq. (8). With different construction

of $\widetilde{C_{xx}}, \widetilde{C_{yy}}$ and $\widetilde{C_{xy}}$ the framework leads to different algorithm, *e.g.*, CCA, KCCA, locality preserving CCA (LPCCA) [12], SCCA and SSCCA. We will briefly list the construction of $\widetilde{C_{xx}}, \widetilde{C_{yy}}$ and $\widetilde{C_{xy}}$ for these algorithms as follows.

CCA:

In original CCA algorithm, the construction the three covariance matrix are directly based on the original mean-normalized vector X and Y, so we obtain:

$$\widetilde{C_{xx}} = XX^T, \widetilde{C_{yy}} = YY^T, \widetilde{C_{xy}} = XY^T.$$

KCCA:

Suppose there are two implicit nonlinear mapping, Φ: $x \mapsto \Phi(x)$ and Ψ: $y \mapsto \Psi(y)$, which project the original data sets of X and Y to corresponding feature space $\mathbb{F}_x$ and $\mathbb{F}_y$, then the CCA is performed in the feature space $\mathbb{F}_x$ and $\mathbb{F}_y$. Let $\Phi(X) = [\Phi(x_1), \ldots, \Phi(x_n)]$ and $\Psi(Y) = [\Psi(y_1), \ldots, \Psi(y_n)]$, using the kernel trick, we obtain $(K_x)_{ij} = \Phi(x_i)^T\Phi(x_j)$ and $(K_y)_{ij} = \Psi(y_i)^T\Psi(y_j)$, then according to the dual representation theorem, we obtain:

$$\widetilde{C_{xx}} = K_x^2, \widetilde{C_{yy}} = K_y^2, \widetilde{C_{xy}} = K_xK_y.$$

LPCCA:

Let $LN(x_i)$ and $LN(y_i)$ denote the samples set of the local neighbor of x_i and y_i, respectively, we define $S_x = \left\{S_{ij}^x\right\}_{i,j=1}^n$ and $S_y = \left\{S_{ij}^y\right\}_{i,j=1}^n$, where

$$S_{ij}^x = \begin{cases} \exp(-\left\|x_i - x_j\right\|^2 / t_x), \text{ if } x_j \in LN(x_i) \text{ or } x_i \in LN(x_j) \\ 0, \qquad\qquad \text{otherwise} \end{cases},$$

$$S_{ij}^y = \begin{cases} \exp(-\left\|y_i - y_j\right\|^2 / t_y), \text{ if } y_j \in LN(y_i) \text{ or } y_i \in LN(y_j) \\ 0, \qquad\qquad \text{otherwise} \end{cases}.$$

Then the construction of $\widetilde{C_{xx}}, \widetilde{C_{yy}}$ and $\widetilde{C_{xy}}$ can be expressed as

$$\widetilde{C_{xx}} = XS_{xx}X^T, \widetilde{C_{yy}} = YS_{yy}Y^T, \widetilde{C_{xy}} = XS_{xy}Y^T$$

where $S_{xx} = D_{xx} - S_x \circ S_x$, $S_{yy} = D_{yy} - S_y \circ S_y$, $S_{xy} = D_{xy} - S_x \circ S_y$, the symbol $\circ$ denotes an operator on two matrices A and B with the same size such that $(A \circ B)_{ij} = A_{ij}B_{ij}$. D_{xx}(D_{yy}, D_{xy}) is a diagonal matrix whose entries are row (or column, due to symmetry) sum of the matrix $S_x \circ S_x$($S_y \circ S_y$, $S_x \circ S_y$).

SCCA:

As defined in Eq. (13), the supervised matrix S, which makes use of the class information of samples, is block-diagonal, and the elements lie on the block-diagonal of S are 1, the others are 0. Then we have the following expressions (please see section 3.1 for more details),

$$\widetilde{C_{xx}} = XSX^T, \widetilde{C_{yy}} = YSY^T, \widetilde{C_{xy}} = XSY^T.$$

SSCCA:

Incorporating the class information and local neighbor information of samples, the construction of $\widetilde{C_{xx}}, \widetilde{C_{yy}}$ and $\widetilde{C_{xy}}$ can be expressed as (please see section 3.2 for more details)

$$\widetilde{C_{xx}} = XD_{xx}X^T, \widetilde{C_{yy}} = YD_{yy}Y^T, \widetilde{C_{xy}} = XK_{xy}Y^T.$$

4.2 Correlation Feature Extraction Power

From the framework of CCA, described in Eq. (8), we can see that all the derivate algorithms of CCA try to maximize the projection of the correlation matrix on pairs of basis vector ω_x and ω_y. In other words, they find pairs of basis vector ω_x and ω_y by maximizing the canonical correlation coefficients. It is well-known that the correlation of the two sets of variables can be studied through a few pairs of canonical variates which are corresponding to the first few canonical correlation coefficients in the framework of CCA. Therefore, the canonical correlation coefficient $\rho(x, y, \omega_x, \omega_y)$ reflects the correlation feature extraction power (CFEP) of different algorithms.

In the CCA algorithm, the eigenvalues in Eq. (3) are just the canonical correlation coefficients. Therefore, the eigenvalues of CCA reflect the CFEP of CCA. In our SSCCA algorithm, the eigenvalues are also the canonical correlation coefficients, which reflect the CFEP of SSCCA too. Following this basic idea, we compare the eigenvalues of CCA and SSCCA to evaluate their CFEP, the result is shown in Fig. 1.

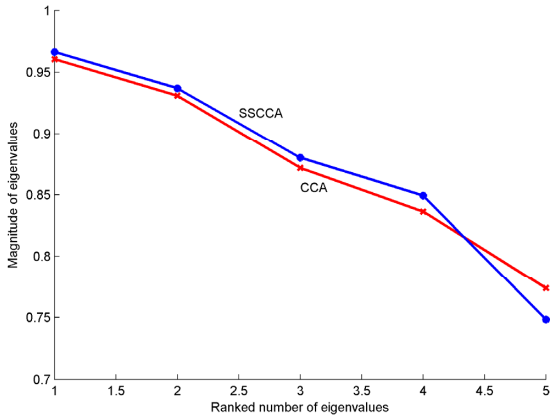

Fig. 1. Eigenvalues of CCA and SSCCA. The abscissa is the ranked number of eigenvalue in descending order and the ordinate is the magnitude of eigenvalue.

The data set used for this study is the JAFFE database (please see section 5.1 for more details). As shown in Fig. 1, the first few eigenvalues of SSCCA are bigger than those of CCA which indicates that SSCCA can have more CFEP than CCA. For classification, the CFEP is directly related to the discriminating power, so we expect that SSCCA can obtain better performance than CCA on the FER problem.

5. EXPERIMENTAL RESULT

In this section, we will evaluate the validity of the SSCCA algorithm for FER on JAFFE and Cohn-Kanade expression databases. The two groups of features which we use for combination are Fisherface feature [4] and Laplacianface feature [5]. The fusion strategy (5) is adopted to extract the fusion features for classification.

Firstly, the SSCCA algorithm is compared with Laplacianface and Fisherface to show the advantages of the combined feature to single feature. Then the effectiveness of SSCCA algorithm is shown by comparing with CCA, KCCA, LPCCA and SCCA. In addition, the nearest neighbor classifier is used throughout the following experiments for its simplicity.

5.1 FER on the JAFFE Database

The JAFFE database [22] consists of 213 images from 10 individuals of Japanese female, covering seven categories of basic facial expressions (neutral, anger, disgust, fear, happiness, sadness and surprise). The original images all have the same size of 256 × 256 pixels with 256-level gray scale. The images are cropped automatically to make two eyes align at the same position and then resized to 100 × 100 pixels. Some cropped images are shown in Fig. 2.

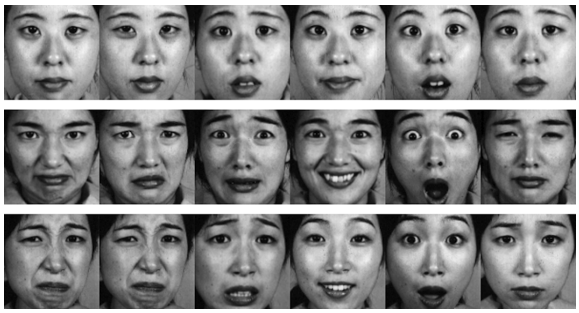

Fig. 2. Cropped images from the JAFFE facial expression database.

Except for the neutral expression, we choose 21 images from each of the remaining six facial expressions. We adopt the leave-one-out cross-validation technique to verify our algorithm. One image is chosen from each expression for testing, while the left twenties are used for training. This should be repeated 21 times, and the average recognition accuracy is taken as the final recognition accuracy.

The single features based on Laplacianface and Fisherface are first calculated respectively, then the combined feature is extracted using SSCCA. The top recognition accuracies with their corresponding dimensions of the three methods are shown in Table 1. Note that, both the Laplacianface method and Fisherface method should adopt the PCA algorithm to reduce the dimension of the feature space to avoid the singular problem, and then the corresponding algorithm, *i.e.* LPP or LDA, is followed to extract the facial expression features respectively.

In the unified framework of CCA, when different correlation criterion is adopted, the classification performance of the combined feature differs from each other. We compare the performance of different forms of CCA, *i.e.* CCA, KCCA, LPCCA, SCCA and SSCCA, the recognition accuracies and the corresponding dimensions are shown in Table 2. The Gaussian function $k(x, y) = \exp(-\|x - y\|^2/t)$ is used to define the kernel matrix in KCCA and the affinity matrix in SSCCA and LPCCA. The parameters t_x and t_y in SSCCA are chosen to make sure that S_x and S_y are on the same magnitude. In our experiment, the parameters t_x and t_y in KCCA, LPCCA and SSCCA are adjusted for the best performance respectively, and the results are also shown in Table 2.

Table 1. Experimental results of Laplacianface, Fisherface and SSCCA on JAFFE.

	Laplacianface	Fisherface	SSCCA
Accuracy (%)	92.06	92.86	96.03
Dimension (d)	7	5	5

Table 2. Experimental results of different forms of CCA on JAFFE.

	CCA	KCCA (t_x = 10, t_y = 10)	LPCCA (t_x = 1e6, t_y = 1e6)	SCCA	SSCCA (t_x = 10, t_y = 10)
Accuracy (%)	93.65	94.44	93.65	94.44	96.03
Dimension (d)	5	62	5	5	5

As shown in Table 1, the recognition accuracy of the combined feature extracted from SSCCA is greatly improved than those of the single feature based methods. After combing two single features, the extracted feature contains more effective discriminative information, so the recognition accuracy is greatly improved. In addition, the dimension of our method is the minimal dimension, that is min(p, q), of the two single feature combined, which indicates that we can obtain more effective facial expression representation with the minimal dimension of the single feature.

From Table 2, we can learn that the recognition accuracies of SCCA and SSCCA are higher than those of CCA, KCCA and LPCCA, which shows that the supervised extension of CCA is more suitable for classification. This can be attributed to the introducing of class information into the computation of the correlation projection vectors pairs (ω_x, ω_y). Furthermore, our experiment results in Table 2 also show that SSCCA can achieve better performance than SCCA. This is can be explained that the incorporating of local structure information of the samples in the same class to construct the within-set covariance matrices $\widetilde{C_{xx}}$ and $\widetilde{C_{yy}}$ and the between-set covariance matrix $\widetilde{C_{xy}}$ makes the combined feature more effective to reveal the intrinsic characters of different facial expressions.

5.2 FER on the Cohn-Kanade Database

The Cohn-Kanade (CK) database [23] consists of approximately 500 image sequences from 100 subjects. Each image sequence displays distinct facial expressions, starting from neutral expression and ending with the peak of the expression. For each

expression of a subject, the last eight frames in the image sequences are selected. The images are cropped automatically to make two eyes align at the same position and then resized to 64 × 64 pixels. Some cropped images are shown in Fig. 3.

Fig. 3. Cropped images from the CK facial expression database.

Except for the neutral expression, we choose 160 images from each of the remaining six facial expressions, *i.e.* anger, disgust, fear, happiness, sadness and surprise, for training and testing. For each facial expression, we randomly select k(k = 10, 20, …, 80) images for training and the remaining 160-k images are used for testing. We repeat the experiment 10 times, and the average recognition accuracies are taken as the final recognition accuracies.

Firstly, we compare the recognition performance of Laplacianface, Fisherface and SSCCA, the average recognition accuracies are shown in Fig. 4. Then we compare the performance of different forms of CCA, the results are shown in Fig. 5. Finally, we list the dimensions of feature corresponding to the top recognition accuracies of different forms CCA by setting k = 30 in Table 3.

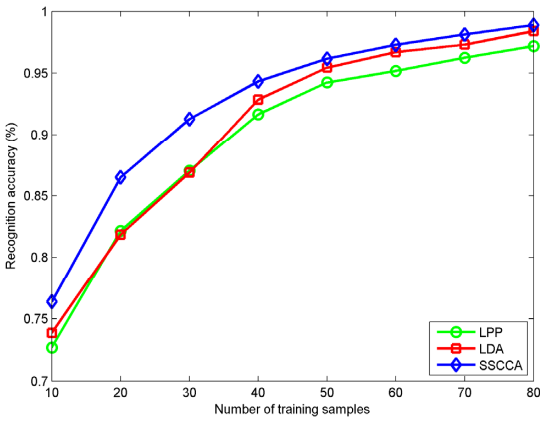

Fig. 4. Average recognition accuracies of Laplacianface, Fisherface and SSCCA.

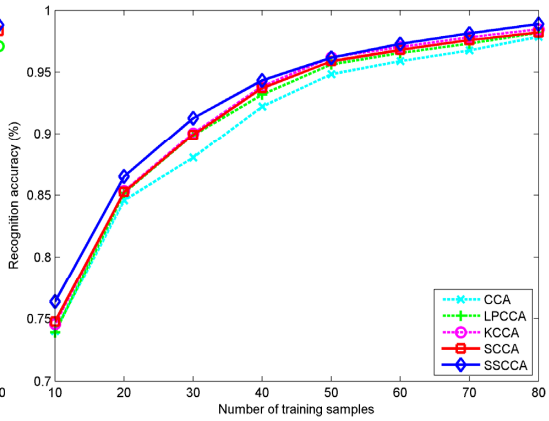

Fig. 5. Average recognition accuracies of different forms of CCA.

Table 3. Experimental results of different forms of CCA on CK (k=30).

	CCA	KCCA (t_x = 1e3, t_y =1e3)	LPCCA (t_x = 1e6, t_y = 1e6)	SCCA	SSCCA (t_x = 10, t_y = 10)
Accuracy (%)	88.03	89.87	90.03	89.90	91.26
Dimension (d)	5	77	5	5	5

Generally, the performances of all these algorithms vary with the size of the training dataset. As is shown in Fig. 4, the recognition accuracies of our method are consistently higher than those of the single feature based methods. From Fig. 5 and Table 3, we can see that SSCCA outperforms other forms of CCA with higher recognition accuracy and lower feature dimensionality. With the increase in the number of training samples, the KCCA and SCCA methods perform comparatively to SSCCA. Furthermore, the advantage of SSCCA is more obvious when there are less training samples. As a matter of fact, we may not have sufficient training samples in the real world's applications. And the superiority of our algorithm is obvious when there are less training samples, so it can be more valuable for real world's applications.

6. CONCLUSION

Inspired by the successful application of spectral graph theory in classification, we proposed a novel supervised feature fusion method called supervised spectral canonical correlation analysis (SSCCA) in this paper. In SSCCA, the affinity matrix, which incorporates the class information and local structure information of the data points, is used as the supervised matrix, and then the ratio of the between-set covariance and the within-set covariance is maximized to seek pairs of projection (ω_x, ω_y). SSCCA utilizes the spectral of covariance matrices (within-set covariance matrices and between-set covariance) to obtain a new combined feature, which means it can not only extract the effective information of each single feature, but also eliminate the redundant information within the features, so SSCCA is superior to single feature based method. Further analysis shows that the feature extracted by SSCCA has discriminative information, and it can reveal the intrinsic nonlinear manifold structure hidden in the original data, which implies that SSCCA is suitable for nonlinear recognition problems. Furthermore, a unified framework of CCA is proposed to offer an effective methodology for non-empirical structural comparison of different forms of CCA as well as providing a way to extend the CCA algorithm. The correlation feature extraction power is also introduced to evaluate the performance of different forms of CCA.

The experiments on the databases of JAFFE and CK validate the effectiveness of our method. From the experiments results we can see that SSCCA not only is superior to the single feature based method, but also outperforms other forms of CCA in terms of the recognition performance on the FER problem. Furthermore, it should be noticed that when the samples in different channels are given with their class labels, the application of SSCCA can be extended to various fields such as image retrieval, pattern recognition and other fields.

REFERENCES

1. B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, Vol. 36, 2003, pp. 259-275.
2. M. Pantie and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 1424-1445.
3. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, Vol. 3, 1991, pp. 72-86.
4. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, 1997, pp. 711-720.
5. X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, 2005, pp. 328-340.
6. I. Kotsia, S. Zafeiriou, and I. Pitas, "Texture and shape information fusion for facial expression and facial action unit recognition," *Pattern Recognition*, Vol. 41, 2008, pp. 833-851.
7. W. M. Zheng, X. Y. Zhou, C. R. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Transactions on Neural Networks*, Vol. 17, 2006, pp. 33-238.
8. J. Yang, J. Y. Yang, D. Zhang, and J. F. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern Recognition*, Vol. 36, 2003, pp. 1369-1381.
9. Q. S. Sun, S. G. Zeng, Y. Liu, P. A. Heng, and D. S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, Vol. 38, 2005, pp. 2437-2448.
10. P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *International Journal of Neural System*, Vol. 10, 2002, pp. 365-377.
11. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
12. T. K. Sun and S. C. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, Vol. 25, 2007, pp. 531-543.
13. N. E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of SVM kernels," *Pattern Recognition*, Vol. 38, 2005, pp. 1733-1745.
14. Y. Chang, C. Hu, and M. Turk, "Manifold of facial expression," in *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 28-35.
15. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, Vol. 290, 2000, pp. 2323-2326.
16. L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, Vol. 4, 2003, pp. 119-155.
17. F. Wang, J. Wang, C. Zhang, and J. Kwok, "Face recognition using spectral features," *Pattern Recognition*, Vol. 40, 2007, pp. 2786-2797.
18. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, Vol. 22, 2000, pp. 888-905.
19. U. von Luxburg, "A tutorial on spectral clustering," Technical Report, TR-149, Max Planck Institute for Biological Cybernetics, 2006.
20. S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral learning," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2003, pp. 561-566.
21. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceeding of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200-205.
22. T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceeding of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46-53.

Song Guo (郭松) received the B.S. degree in biomedical engineering from Beijing Jiaotong University, Beijing in 2007. He is currently a Ph.D. student in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include pattern recognition, image processing and machine learning.

Qiuqi Ruan (阮秋琦) received the B.S. and M.S. degrees from Northern Jiaotong University in 1969 and 1981 respectively. From January 1987 to May 1990, he was a visiting scholar in the University of Pittsburgh, and the University of Cincinnati. Subsequently, he has been a Visiting Professor in USA for several times. He has published 4 books and more than 100 papers, and achieved a national patent. Now he is a Professor, doctorate supervisor in Institute of Information Science, Beijing Jiaotong University, Beijing, China. He is a senior member of IEEE. His main research interests include digital signal processing, computer vision, pattern recognition, and virtual reality *etc.*

Zhan Wang (王占) received the B.S. degree in information and computer science from Tai Yuan University of Science and Technology in 2004, and M.S. degree in Applied Mathematics from Beijing Jiaotong University in 2007. He is currently a Ph.D. student in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include image processing, pattern recognition and machine learning.

Shuai Liu (刘帅) received the B.S. degree in Computer Science from Beijing Jiaotong University, P.R. China in 2006. He is currently a Ph.D. student in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include pattern recognition, image processing and machine learning.