Letters

Facial Expression Recognition Using Kernel Canonical Correlation Analysis (KCCA)

Wenming Zheng, Xiaoyan Zhou, Cairong Zou, *Member, IEEE*, and Li Zhao

***Abstract*—In this correspondence, we address the facial expression recognition problem using kernel canonical correlation analysis (KCCA). Following the method proposed by Lyons *et al.* [7] and Zhang *et al.* [8], we manually locate 34 landmark points from each facial image and then convert these geometric points into a labeled graph (LG) [7] vector using the Gabor wavelet transformation method to represent the facial features. On the other hand, for each training facial image, the semantic ratings describing the basic expressions are combined into a six-dimensional semantic expression vector. Learning the correlation between the LG vector and the semantic expression vector is performed by KCCA. According to this correlation, we estimate the associated semantic expression vector of a given test image and then perform the expression classification according to this estimated semantic expression vector. Moreover, we also propose an improved KCCA algorithm to tackle the singularity problem of the Gram matrix. The experimental results on the Japanese female facial expression database and the Ekman's "Pictures of Facial Affect" database illustrate the effectiveness of the proposed method.**

***Index Terms*—Facial expression recognition (FER), generalized discriminant analysis (GDA), kernel canonical correlation analysis (KCCA), kernel method.**

I. Introduction

Automatic facial expression recognition (FER) could be traced back to the preliminary work of Suwa *et al.* [15] in 1978 and gained much popularity starting with the pioneering work of Mase and Pentland [16] in the 1990s. More recently, facial expression analysis has become a very hot research topic in computer vision and pattern recognition, and various approaches have been proposed to this goal. For literature surveys, see [12]–[14]. The contribution in this correspondence is two fold: 1) we present an effective facial expression analysis and recognition method based on kernel canonical correlation analysis (KCCA), and 2) we propose an improved algorithm for KCCA to tackle the singularity problem of the Gram matrix.

Following the work done by Lyons *et al.* [6], [7] and Zhang *et al.* [8], and being motivated by the kernel-based machine learning algorithms [18], we first manually locate 34 landmark points from each facial image and then convert these geometric points into a labeled graph (LG) [7] vector using the Gabor wavelet transformation method [20], [21] to represent the facial features. On the other hand, a semantic expression vector consisting of the semantic ratings of each training facial image is used as the semantic features. Learning the correlation between the LG vector and the semantic expression vector is performed by KCCA. According to this correlation, we estimate the associated semantic expression vector of a given test image and then perform the expression classification according to this estimated semantic expression vector.

Canonical correlation analysis (CCA) is a powerful method to correlate the linear relation between two multidimensional variables. However, if there is nonlinear correlation between the two variables, linear CCA may not correctly correlate this relationship. KCCA is a nonlinear extension of CCA via the "kernel trick" to overcome this drawback [10], [11]. Although the idea of kernelizing CCA is not new, all the previous algorithms of KCCA [2], [3] tackle the singularity problem of the Gram matrix by simply adding a regularization to the Gram matrix such that the Gram matrix becomes invertible [2], [3]. In this correspondence, we propose an improved KCCA algorithm based on eigenvalue decomposition approach rather than the regularization method to overcome this problem.

The remainder of this correspondence is organized as follows: we derive the KCCA formulations in Section II, and the facial expression classification formulations based on KCCA in Section III. Experiments are conducted in Section IV, and conclusion are given in Section V.

II. KCCA

Given two centered random multivariables $\mathbf{x} \in \mathbf{R}^{n_x}$ and $\mathbf{y} \in \mathbf{R}^{n_y}$, the goal of CCA is to find a pair of directions $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ such that the correlation $\rho(\mathbf{x}, \mathbf{y})$ between the two projections $\boldsymbol{\omega}_x^T \mathbf{x}$ and $\boldsymbol{\omega}_y^T \mathbf{y}$ is maximized [1], where

$$\begin{aligned}\rho(\mathbf{x}, \mathbf{y}; \boldsymbol{\omega}_x, \boldsymbol{\omega}_y) &= \frac{E\left\{\boldsymbol{\omega}_x^T \mathbf{x}\mathbf{y}^T \boldsymbol{\omega}_y\right\}}{\sqrt{E\left\{\boldsymbol{\omega}_x^T \mathbf{x}\mathbf{x}^T \boldsymbol{\omega}_x\right\}}\sqrt{E\left\{\boldsymbol{\omega}_y^T \mathbf{y}\mathbf{y}^T \boldsymbol{\omega}_y\right\}}} \\ &= \frac{\boldsymbol{\omega}_x^T E\{\mathbf{x}\mathbf{y}^T\}\boldsymbol{\omega}_y}{\sqrt{\boldsymbol{\omega}_x^T E\{\mathbf{x}\mathbf{x}^T\}\boldsymbol{\omega}_x}\sqrt{\boldsymbol{\omega}_y^T E\{\mathbf{y}\mathbf{y}^T\}\boldsymbol{\omega}_y}}. \qquad (1)\end{aligned}$$

Suppose that $\{\mathbf{x}_i\}_{i=1,\ldots,N}$ and $\{\mathbf{y}_i\}_{i=1,\ldots,N}$ are N observations of $\mathbf{x}$ and $\mathbf{y}$ respectively. Then the goal of CCA is equivalent to finding $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ that maximize

$$\rho(\mathbf{x}, \mathbf{y}; \boldsymbol{\omega}_x, \boldsymbol{\omega}_y) = \frac{\boldsymbol{\omega}_x^T \mathbf{X}\mathbf{Y}^T \boldsymbol{\omega}_y}{\sqrt{\boldsymbol{\omega}_x^T \mathbf{X}\mathbf{X}^T \boldsymbol{\omega}_x}\sqrt{\boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \boldsymbol{\omega}_y}} \qquad (2)$$

where

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_N], \quad \mathbf{Y} = [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \cdots \quad \mathbf{y}_N].$$

Let $\mathbf{x}$ be mapped into a Hilbert space $\mathbf{F}$ through a nonlinear mapping Φ

$$\Phi : \mathbf{R}^{n_x} \to \mathbf{F}, \quad \mathbf{x} \to \Phi(\mathbf{x}).$$

Also, we suppose that $\Phi(\mathbf{x})$ is centered in $\mathbf{F}$. For a method to center $\Phi(\mathbf{x})$, see [11]. From (2), we obtain that the correlation function in $\mathbf{F}$ can be formulated as

$$\rho(\Phi(\mathbf{x}), \mathbf{y}; \boldsymbol{\omega}_{\Phi(\mathbf{x})}, \boldsymbol{\omega}_y) = \frac{\boldsymbol{\omega}_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})\mathbf{Y}^T \boldsymbol{\omega}_y}{\sqrt{\boldsymbol{\omega}_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})(\Phi(\mathbf{X}))^T \boldsymbol{\omega}_{\Phi(\mathbf{x})}}\sqrt{\boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \boldsymbol{\omega}_y}}.$$

Manuscript received March 16, 2004; revised November 6, 2004. This work was supported in part by the Jiangsu Natural Science Foundations under Grants BK2005407 and BK2005122 and in part by the Natural Science Foundations under Grant 60503023.

W. Zheng is with the Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: wenming_zheng@seu.edu.cn).

X. Zhou is with the Department of Electrics Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu 210044, China (e-mail: xiaoyan_zhou@nuist.edu.cn).

C. Zou and L. Zhao are with the Engieneering Research Center of Information Processing and Application, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: cairong@seu.edu.cn; zhaoli@seu.edu.cn).

Digital Object Identifier 10.1109/TNN.2005.860849

Therefore, the goal of KCCA is equivalent to finding $\omega_{\Phi(x)}$ and ω_y that maximize $\omega_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})\mathbf{Y}^T\omega_y$ under the constraints

$$\boldsymbol{\omega}_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})(\Phi(\mathbf{X}))^T \omega_{\Phi(\mathbf{x})} = \boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \omega_y = 1 \tag{3}$$

where

$$\Phi(\mathbf{X}) = [\Phi(\mathbf{x}_1) \quad \Phi(\mathbf{x}_2) \quad \cdots \quad \Phi(\mathbf{x}_N)]. \tag{4}$$

According to [17], there exists direction $\boldsymbol{\alpha}_{\Phi(\mathbf{x})}$ such that

$$\omega_{\Phi(\mathbf{x})} = (\Phi(\mathbf{X}))\alpha_{\Phi(\mathbf{x})}. \tag{5}$$

Assume that $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function that can be expressed as the dot product form on the Hilbert space $\mathbf{F}$

$$(k)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = (\Phi(\mathbf{x}_i))^T \Phi(\mathbf{x}_j) \tag{6}$$

where $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ stands for the dot product of $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$.

From (4) and (6), we obtain that the $N \times N$ matrix $\mathbf{K} = ((k)_{ij})_{i=1,\ldots,N; j=1,\ldots,N}$ (also known as the Gram matrix) can be written as

$$\mathbf{K} = (\Phi(\mathbf{X}))^T \Phi(\mathbf{X}). \tag{7}$$

From (5), (3), and (7), we obtain

$$\omega_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})\mathbf{Y}^T \boldsymbol{\omega}_y = \boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{Y}^T \omega_y \tag{8}$$

$$\begin{aligned} \boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \boldsymbol{\omega}_y &= \boldsymbol{\omega}_{\Phi(\mathbf{x})}^T \Phi(\mathbf{X})(\Phi(\mathbf{X}))^T \omega_{\Phi(\mathbf{x})} \\ &= \boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = 1. \end{aligned} \tag{9}$$

Thus, solving the KCCA problem is equivalent to finding $\boldsymbol{\alpha}_{\Phi(x)}$ and ω_y that maximize $\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{Y}^T \boldsymbol{\omega}_y$ under the constraints

$$\boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \boldsymbol{\omega}_y = \boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = 1.$$

The corresponding Lagrangian is

$$\begin{aligned} &L\left(\boldsymbol{\alpha}_{\Phi(\mathbf{x})}, \omega_y, \lambda, \mu\right) \\ &= \boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{Y}^T \omega_y - \lambda \left(\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} - 1\right) \Big/ 2 \\ &- \mu\left(\boldsymbol{\omega}_y^T \mathbf{Y}\mathbf{Y}^T \boldsymbol{\omega}_y - 1\right) \Big/ 2. \end{aligned}$$

Taking derivatives with respect to $\boldsymbol{\alpha}_{\Phi(\mathbf{x})}$ and ω_y and set to zeros, we obtain

$$\frac{\partial L}{\partial \boldsymbol{\alpha}_{\Phi(\mathbf{x})}} = \mathbf{K}\mathbf{Y}^T \boldsymbol{\omega}_y - \lambda \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = \mathbf{0} \tag{10}$$

$$\frac{\partial L}{\partial \omega_y} = \mathbf{Y}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} - \mu \mathbf{Y}\mathbf{Y}^T \omega_y = \mathbf{0}. \tag{11}$$

From (9)–(11), we obtain

$$\mu = \lambda, \omega_y = \frac{(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{K}}{\mu} \boldsymbol{\alpha}_{\Phi(\mathbf{x})} \tag{12}$$

$$\mathbf{K}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = \lambda^2 \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})}. \tag{13}$$

If the Gram matrix $\mathbf{K}$ has full rank, then (13) can be rewritten as

$$\mathbf{K}^{-1}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = \lambda^2 \boldsymbol{\alpha}_{\Phi(\mathbf{x})}. \tag{14}$$

However, in most cases, $\mathbf{K}$ is often singular because it is a centered Gram matrix [2]. To overcome this problem, one may add a regularization matrix $N_{\mathbf{K}}\mathbf{I}$ to $\mathbf{K}$ and then resolve the following generalized eigenequation (14), [2], [3], [5]

$$(\mathbf{K} + N_{\mathbf{K}}\mathbf{I})^{-1}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})} = \lambda^2 \boldsymbol{\alpha}_{\Phi(\mathbf{x})} \tag{15}$$

where $\mathbf{I}$ is the identity matrix and $N_{\mathbf{K}}$ is a small constant. The drawback of this approach is that the selection of the regularization to the Gram matrix could influence the performance of KCCA. In what follows, we will propose an improved algorithm for KCCA based on eigenvalue decomposition to avoid the regularization to the Gram matrix. In fact, the largest eigenvalue of (13) gives the maximum of the following Rayleigh quotient [4]

$$\lambda^2 = \frac{\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})}}{\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})}}. \tag{16}$$

Let $\mathbf{W} = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}$, then (16) can be rewritten as

$$\lambda^2 = \frac{\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{W}\mathbf{K}\alpha_{\Phi(\mathbf{x})}}{\boldsymbol{\alpha}_{\Phi(\mathbf{x})}^T \mathbf{K}\mathbf{K}\boldsymbol{\alpha}_{\Phi(\mathbf{x})}}. \tag{17}$$

Thus, solving KCCA is equivalent to finding $\boldsymbol{\alpha}_{\Phi(\mathbf{x})}$ that maximizes the Rayleigh quotient in (17), which is equivalent to solving the same optimal problem as that of generalized discriminant analysis (GDA) [4]. Therefore, we can adopt the same eigenvalue decomposition approach as that used in GDA for the Gram matrix to system (17). Moreover, note that there exists the degenerate eigenvalue problem in solve system (17), we propose to use the method used in the modified generalized discriminant analysis (MGDA) [9] algorithm to tackle this problem.

III. Pattern Classification Using KCCA

Let $\{(\omega_{\Phi(\mathbf{x})}^i, \omega_{\mathbf{y}}^i)\}_{i=1}^t$ be the t pair directions of KCCA, and $\rho_1, \ldots, \rho_t$ the t corresponding correlation values. Let a_i and b_i be the projections of the variables $\Phi(\mathbf{x})$ and $\mathbf{y}$ onto the projection vectors $\boldsymbol{\omega}_{\Phi(\mathbf{x})}^i$ and $\boldsymbol{\omega}_{\mathbf{y}}^i$, respectively. Thus, we get

$$a_i = \left(\boldsymbol{\omega}_{\Phi(\mathbf{x})}^i\right)^T \Phi(\mathbf{x}), \quad b_i = \left(\boldsymbol{\omega}_y^i\right)^T \mathbf{y}(i = 1, \ldots, t).$$

Suppose that $\rho_i \approx 1$. Then, we can obtain that a_i and b_i are proximately linearly correlated; thus, we can assume that there exists coefficient k_i and constant ε_i such that

$$b_i = k_i a_i + \varepsilon_i (i = 1, \ldots, t). \tag{18}$$

Note that both b_i and a_i are centered projections. Thus, according to regression analysis [19], the coefficients k_i and the constant ε_i can be calculated using the observations of $\Phi(\mathbf{x})$ and $\mathbf{y}$, as follows:

$$\begin{aligned} k_i = &\sum_{j=1}^N \left[\left(\boldsymbol{\omega}_{\Phi(\mathbf{x})}^i\right)^T \Phi(\mathbf{x}_j) \left(\boldsymbol{\omega}_{\mathbf{y}}^i\right)^T \mathbf{y}_j\right] \Bigg/ \\ &\sum_{j=1}^N \left[\left(\boldsymbol{\omega}_{\Phi(\mathbf{x})}^i\right)^T \Phi(\mathbf{x}_j)\right]^2 \end{aligned} \tag{19}$$

$$\begin{aligned} \varepsilon_i = &\frac{1}{N} \sum_{j=1}^N \left(\boldsymbol{\omega}_{\mathbf{y}}^i\right)^T \mathbf{y}_j \\ &- \frac{k_i}{N} \sum_{j=1}^N \left(\boldsymbol{\omega}_{\Phi(\mathbf{x})}^i\right)^T \Phi(\mathbf{x}_j) = 0. \end{aligned} \tag{20}$$

Let $\Phi(\mathbf{x}_{\text{test}})$ be a test sample and $\mathbf{y}_{\text{test}}$ the corresponding observation of $\mathbf{y}$. Let $\mathbf{a}_{\text{test}}$ be the projection of $\Phi(\mathbf{x}_{\text{test}})$ onto the directions $\omega_{\Phi(\mathbf{x})}^1, \ldots, \omega_{\Phi(\mathbf{x})}^t$, and $\mathbf{b}_{\text{test}}$ the projection of $\mathbf{y}_{\text{test}}$ onto the directions $\omega_y^1, \ldots, \omega_y^t$. Then, we get

$$\begin{aligned} \mathbf{a}_{\text{test}} &= \left[a_{\text{test}}^1, \ldots, a_{\text{test}}^t\right] \\ &= \left[\boldsymbol{\omega}_{\Phi(\mathbf{x})}^1, \ldots, \boldsymbol{\omega}_{\Phi(\mathbf{x})}^t\right]^T \Phi(\mathbf{x}_{\text{test}}) = \mathbf{P}_{\mathbf{x}}^T \mathbf{K}_{\text{test}} \end{aligned} \tag{21}$$

$$\begin{aligned} \mathbf{b}_{\text{test}} &= \left[b_{\text{test}}^1, \ldots, b_{\text{test}}^t\right] \\ &= \left[\boldsymbol{\omega}_{\mathbf{y}}^1, \ldots, \boldsymbol{\omega}_{\mathbf{y}}^t\right]^T \mathbf{y}_{\text{test}} = \mathbf{P}_{\mathbf{y}}^T \mathbf{y}_{\text{test}} \end{aligned} \tag{22}$$

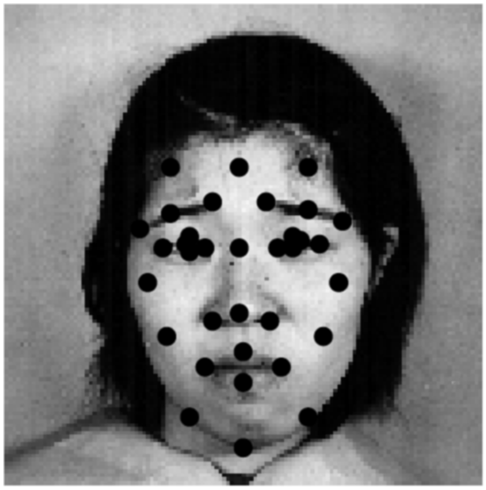

Fig. 1. Example of the 34 landmark points selected from a facial image.

where

$$\mathbf{P_x} = \left[\boldsymbol{\alpha}^1_{\Phi(\mathbf{x})}, \ldots, \boldsymbol{\alpha}^t_{\Phi(\mathbf{x})}\right]$$
$$\mathbf{K}_{\text{test}} = (\Phi(\mathbf{X}))^T \Phi(\mathbf{x}_{\text{test}})$$
$$\mathbf{P_y} = \left[\omega^1_{\mathbf{y}}, \ldots, \omega^t_{\mathbf{y}}\right].$$

The projection $\mathbf{y}_{\text{test}}$ in (22) can be solved by

$$\mathbf{y}_{\text{test}} = \left(\mathbf{P_y}\mathbf{P_y}^T\right)^{-1}\mathbf{P_y}\mathbf{b}_{\text{test}}. \tag{23}$$

If $\mathbf{P_y}\mathbf{P_y}^T$ is singular, we add a regularization $\varepsilon\mathbf{I}$ on $\mathbf{P_y}\mathbf{P_y}^T$ such that $\mathbf{P_y}\mathbf{P_y}^T + \varepsilon\mathbf{I}$ is a nonlinear matrix. Then (23) can rewritten by

$$\mathbf{y}_{\text{test}} = \left(\mathbf{P_y}\mathbf{P_y}^T + \varepsilon\mathbf{I}\right)^{-1}\mathbf{P_y}\mathbf{b}_{\text{test}}. \tag{24}$$

Thus, for a given test observation $\mathbf{x}_{\text{test}}$, we can estimate the corresponding observation $\mathbf{y}_{\text{test}}$ using (21), (18), (19), (20) and (24). Let y^i_{test} denote the ith element of $\mathbf{y}_{\text{test}}$, then the index of the most matched expression class of the test image is signed by

$$c^* = \arg\max_i y^i_{\text{test}}. \tag{25}$$

IV. Experiments

We will use the Japanese female facial expression (JAFFE) database [6]–[8] and Ekman's "Pictures of Facial Affect" database [22], respectively, to conduct the FER based on KCCA. In the preprocessing stage, we manually locate 34 landmark points from each facial image by referring to the method in [7]. Fig. 1 illustrates an example of the 34 landmark points.

Similar with Lyons *et al.* [7], after locating the 34 landmark points, we use the Gabor wavelet representation of each facial image at the landmark points to represent the facial features of each image, where all of the wavelet convolution values (magnitudes) at these landmark points are combined into a 1020-dimensional LG vector. The Gabor kernel is defined as follows [7], [20], [21]:

$$\psi_{u,v}(\mathbf{z}) = \frac{\|\mathbf{k}_{u,v}\|^2}{\sigma^2}\exp\left(-\frac{\|\mathbf{k}_{u,v}\|^2\|\mathbf{z}\|^2}{2\sigma^2}\right) \times \left[\exp(i\mathbf{k}_{u,v}\cdot\mathbf{z}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{26}$$

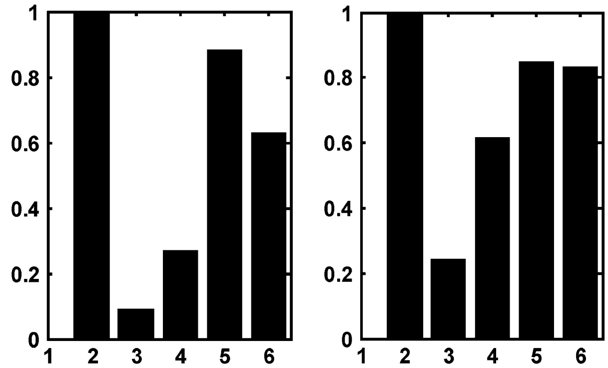

Fig. 2. Semantic ratings (1: happiness; 2: sadness; 3: surprise; 4: anger; 5: disgust; 6: fear) of the facial image in Fig. 1, where the left figure illustrates the semantic ratings from psychologists, whereas the right one illustrates the semantic ratings estimated by the KCCA method.

where u and v represent the orientation and scale of the Gabor kernels, and $\mathbf{k}_{u,v}$ is defined as

$$\mathbf{k}_{u,v} = k_v \exp(i\phi_u) \tag{27}$$

where $k_v = \pi/2^v (v \in \{1, \ldots, 5\})$ and $\phi_u = \pi u/6 (u \in \{0, \ldots, 5\})$.

It is notable that both JAFFE database and Ekman's expression database provide semantic ratings evaluated by psychologists for each facial image. These semantic ratings are used for quantitatively evaluating the six basic expressions (happiness, sadness, surprise, anger, disgust, and fear). In this experiment, the semantic ratings of each image are combined into a six-dimensional semantic expression vector for expression analysis.

Let $\mathbf{x}$ and $\mathbf{y}$, respectively, denote the LG vector and semantic expression vector. The goal of the KCCA-based facial expression analysis method is to estimate the associated value of $\mathbf{y}$ for a given observation of $\mathbf{x}$ using (24) and then perform the facial expression classification task using (25). We use the "leave-one-image-out" and "leave-one-subject-out" cross validation strategies, respectively, to perform the experiment. In the "leave-one-image-out" strategy, one facial image is used as the testing data and the other ones as the training data. This is repeated for all of the possible trials and the recognition results are averaged as the final recognition rate. In the "leave-one-subject-out" strategy, the images belonging to one subject are used as the testing data and the remainder ones as the training data. This is also repeated for all of the possible trials until all the subjects are used as the testing data. The experiment results are also averaged as the final recognition rate.

A. FER on JAFFE Facial Expression Database

We use the JAFFE database to test the performance of the proposed method in this experiment. There are 213 images covering seven categories of facial expressions (neutral, happiness, sadness, surprise, anger, disgust, and fear) in JAFFE database. Except for the neutral expression images, there are 183 images in this database. The original images are all sized 256×256 pixels with a 256-level gray scale. For a new test facial image, we locate the 34 landmark points and then get the LG vector. The corresponding semantic ratings of each facial image are estimated using (24), and then perform the expression classification according to (25). Fig. 2 illustrates the semantic ratings of the facial image in Fig. 1, where the left figure illustrates the semantic ratings evaluated from psychologists whereas the right one the semantic ratings estimated from our experiment (both figures are normalized into the scope [0, 1]). From Fig. 2, we can see that the

TABLE I
EXPERIMENTAL RESULTS ON JAFFE DATABASE USING THE "LEAVE-ONE-IMAGE-OUT" CROSS VALIDATION WITH SEMANTIC INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
CCA/KCCA (Monomial kernel, $d = 1$)	5	78.14%
KCCA (Monomial kernel, $d = 2$)	5	81.42%
KCCA (Monomial kernel, $d = 3$)	5	80.87%
KCCA (Monomial kernel, $d = 4$)	5	81.97%
KCCA (Gaussian kernel, $\sigma = 1e6$)	5	**85.79%**

TABLE II
EXPERIMENTAL RESULTS ON JAFFE DATABASE USING THE "LEAVE-ONE-SUBJECT-OUT" CROSS VALIDATION WITH SEMANTIC INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
CCA/KCCA (Monomial kernel, $d = 1$)	5	53.55%
KCCA (Monomial kernel, $d = 2$)	5	57.92%
KCCA (Monomial kernel, $d = 3$)	5	60.66%
KCCA (Monomial kernel, $d = 4$)	5	63.93%
KCCA (Gaussian kernel, $\sigma = 21e6$)	5	**74.32%**

right figure approximately coincides with the left one. Based on the semantic ratings in Fig. 2, we can classify this facial image into sadness category. The experimental results based the "leave-one-image-out" and the "leave-one-subject-out" cross validation strategies are shown in Table I and Table II, respectively, where the kernel functions are defined as:

Monomial kernel: $k(x,y) = (\langle x,y\rangle)^d$, where d is the monomial degree; Gaussian kernel: $k(x,y) = \exp(-\|x-y\|^2/\sigma)$, where σ is the parameter of the kernel function.

To further test the performance of the proposed method, we set the values of the semantic expression vector of each training image by referring to the class label information: Assume that $\mathbf{y}_i$ is the semantic expression vector corresponding to the ith training image $\mathbf{x}_i$, and let $\mathbf{y}_{ik}$ denote the $k$thelement of $\mathbf{y}_i$. If $\mathbf{x}_i$ belongs to class j, then we set $\mathbf{y}_{ij} = 1$; else $\mathbf{y}_{ik} = 0$ $(k \neq j)$. In this case, the directions $\omega_{\Phi(x)}$ of KCCA could be equivalent to the discriminant vectors of GDA. For the comparison purpose, we also conduct the same expression classification task using the GDA method, where the nearest neighbor classifier is used for classification. Table III and Table IV show the experimental results.

From Tables I–IV, we can see that the KCCA method achieves much better performance than the CCA method for the FER task. When the class label information of each facial image is used for constructing the semantic expression vector, the recognition rate of KCCA can be achieved as high as 98.36% for the "leave-one-image-out" cross validation experiment and 77.05% for the "leave-one-subject-out" cross validation experiment. Additionally, it is notable that the best recognition results in Tables I and II are, respectively, lower than those in Tables III and IV. The reason could be attributed to the fact that the semantic expression vectors used in Tables III and IV consider the class label information; thus, it will be advantageous when used for the classification task. The drawback of doing this, however, is that it may not quantitatively estimate all the basic expressions from the facial image. Besides, although KCCA and GDA are equivalent when the semantic expression vector of each facial image consists of the class label information in KCCA, our experimental results also show that KCCA may achieve better performance than GDA. The reason could be attributed to the degenerate eigenvalue problem of GDA [9]. Due to the degenerate eigenvalue problem, GDA becomes instable and could not achieve the best recognition result [9]. The improved KCCA algorithm we proposed in this correspondence, however, can overcome this problem and, thus, could obtain better results.

TABLE III
EXPERIMENTAL RESULTS ON JAFFE DATABASE USING THE "LEAVE-ONE-IMAGE-OUT" CROSS VALIDATION WITH CLASS LABEL INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
LDA/GDA (Polynomial kernel, $d = 1$)	5	95.08%
CCA/KCCA (Polynomial kernel, $d = 1$)	5	96.72%
GDA (Polynomial kernel, $d = 2$)	5	95.63%
KCCA (Polynomial kernel, $d = 2$)	5	**98.36%**
GDA (Gaussian kernel, $\sigma = 1e6$)	5	97.81%
KCCA (Gaussian kernel, $\sigma = 1e6$)	5	97.81%

TABLE IV
EXPERIMENTAL RESULTS ON JAFFE DATABASE USING THE "LEAVE-ONE-SUBJECT-OUT" CROSS VALIDATION WITH CLASS LABEL INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
LDA/GDA (Monomial kernel, $d = 1$)	5	64.48%
CCA/KCCA (Monomial kernel, $d = 1$)	5	67.21%
GDA (Monomial kernel, $d = 3$)	5	68.85%
KCCA (Monomial kernel, $d = 3$)	5	70.49%
GDA (Gaussian kernel, $\sigma = 1e6$)	5	72.13%
KCCA (Gaussian kernel, $\sigma = 1e6$)	5	**77.05%**

B. FER on Ekman's Facial Expression Database

In this experiment, we use Ekman's "Pictures of Facial Affect" database [22] to further test the performance of the proposed FER method. This database contains 110 images consisting of six male and eight female subjects. We only choose 96 images (excluding all the 14 neutral images) covering all the 14 persons with six basic facial expressions as the training data. The experiment scheme is designed to be similar with the one in Section A. In the first example of the experiment, we use the semantic ratings provided by the Ekman's database to generate the semantic expression vectors, whereas the second example uses the class label information of each image to generate the semantic expression vectors. The "leave-one-image-out" cross validation and the "leave-one-subject-out" cross validation are also respectively adopted in both examples. The best experimental results of the two examples are shown in Tables V and VI, respectively. Moreover, in Fig. 3 we also show the plot of recognition rates of KCCA and GDA as functions of gaussian kernel parameter σ when the "leave-one-subject-out" cross validation with class label information is used for constructing the semantic expression vector.

V. DISCUSSION AND CONCLUSION

The FER problem based on KCCA has been addressed in this correspondence. The correspondence shows that KCCA is a very effec-

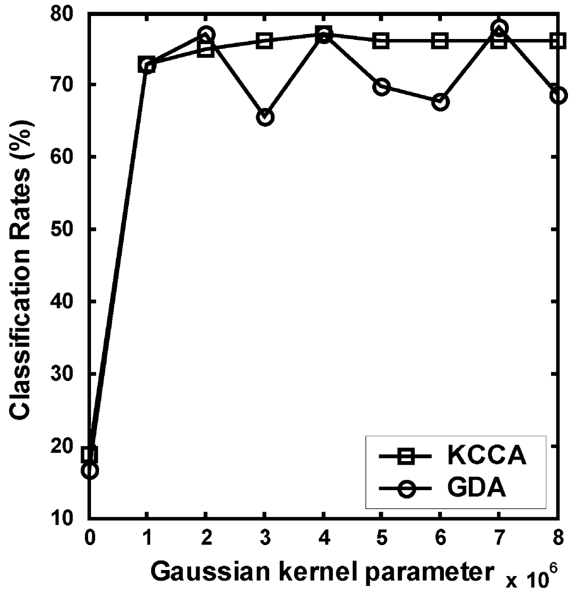

Fig. 3. Recognition rates as functions of the Gaussian kernel parameter σ.

TABLE V
EXPERIMENTAL RESULTS ON EKMAN'S DATABASE WITH SEMANTIC INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
[a] KCCA (Gaussian kernel, $\sigma = 1e7$)	5	81.25%
[b] KCCA (Gaussian kernel, $\sigma = 81e6$)	5	79.17%

Note: [a] Experiment using "leave-one-image-out" strategy
[b] Experiment using "leave-one-subject-out" strategy

TABLE VI
EXPERIMENTAL RESULTS ON EKMAN'S DATABASE WITH CLASS LABEL INFORMATION USED

Method	Number of Canonical Variates	Recognition Rate
[a] GDA (Gaussian kernel, $\sigma = 5e6$)	5	75.00%
[a] KCCA (Gaussian kernel, $\sigma = 5e6$)	5	**77.08%**
[b] GDA (Gaussian kernel, $\sigma = 7e6$)	5	**78.13%**
[b] KCCA (Gaussian kernel, $\sigma = 4e6$)	5	77.08%

Note: [a] Experiment using "leave-one-image-out" strategy
[b] Experiment using "leave-one-subject-out" strategy

tive method for correlating the nonlinear relationship between the facial features and the associated semantic features, which also provide us an effective technique to predict the semantic expression information of a facial image. The experimental results on JAFFE database and Ekman's facial expression database are encouraging. In addition, from the experimental results, we note that the performance of the proposed method is closely related to selection of the semantic expression vectors used for training KCCA. Although the best classification performance in the experiment using JAFFE database is obtained when the class label information is used for constructing the semantic expression vectors, a major drawback of doing this is that these semantic expression vectors convey relatively less basic expression information. Thus, they may not produce better results when used for quantitatively predicting all the six basic expressions. To improve the performance of the semantic ratings-based approach, a natural and better way could resort to obtain more accurate semantic ratings of each facial image to construct the corresponding semantic expression vector. Moreover, in this correspondence we also provide an improved algorithm for KCCA to overcome the singularity problem of the Gram matrix and the degenerate eigenvalue problem of solving the eigenequation of KCCA.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments. They would also like to thank Dr. M. J. Lyons for providing the JAFFE facial expression database and the semantic rating data of the database.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[2] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.
[3] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, 2000.
[4] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.
[5] T. V. Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle, "Kernel canonical correlation analysis and least square support vector machines," in *Proc. Int. Conf. Artificial Neural Networks*, 2001, pp. 384–389.
[6] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition*, Japan, Apr. 1998, pp. 200–205.
[7] M. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.
[8] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition*, Nara, Apr. 1998, pp. 454–459.
[9] W. Zheng, L. Zhao, and C. Zou, "A modified algorithm for generalized discriminant analysis," *Neural Comput.*, vol. 16, no. 6, pp. 1283–1297, 2004.
[10] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
[11] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.
[12] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognit.*, vol. 36, pp. 259–275, 2003.
[13] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[14] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.
[15] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, 1978, pp. 408–410.
[16] K. Mase and A. Pentland, "Recognition of facial expression from optical flow," *Trans. IEICE*, pp. 3474–3483, 1991.
[17] D. R. Hardoon, S. Szedmark, and J. S. Taylor, Canonical Correlation Analysis: An Overview With Application to Learning Methods, Tech. Rep. CSD-TR-03-02, 2003.
[18] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2001.
[19] J. M. Lattin, J. D. Carroll, and P. E. Green, *Analyzing multivariate data*. Pacific Grove, CA: Brooks/Cole, 2003.

[20] L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[21] C. Liu and H. Wechsler, "Independent component analysis of Gabor features for face recognition," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 919–928, Aug. 2003.

[22] P. Ekman and W. V. Friesen, "Pictures of Facial Affect," in *Human Interaction Laboratory*. San Francisco, CA: Univ. California Medical Center, 1976.

Study on Modeling of Multispectral Emissivity and Optimization Algorithm

Chunling Yang, Yong Yu, Dongyang Zhao, and Guoliang Zhao

***Abstract*—Target's spectral emissivity changes variously, and how to obtain target's continuous spectral emissivity is a difficult problem to be well solved nowadays. In this letter, an activation-function-tunable neural network is established, and a multistep searching method which can be used to train the model is proposed. The proposed method can effectively calculate the object's continuous spectral emissivity from the multispectral radiation information. It is a universal method, which can be used to realize on-line emissivity demarcation.**

***Index Terms*—Neural networks, optimization algorithm, spectral emissivity, tunable activation function.**

I. Introduction

The emissivity is not only dependent on the inherence of the object to be measured, but also lies with the physical state of the object's surface, such as surface smoothness, material granularity, temperature, radiant angle, wavelength, and so on. A great of error exists in the emissivity obtained by experienced methods because of the factors listed above. At present, this problem has not been well solved. Classical emissivity measuring method, such as calorimetry, the thermal return reflection method and radiometry [1]–[3] need to install some assistant equipment around the object to be detected and they cannot realize the on-line demarcation. Thus their applications are confined.

Multispectral thermometry is an improved method in recent years for obtaining radiation temperature and emissivity, which can separate emissivity and the real temperature from the object's radiation information. We can use this new method to overcome the above problem. However, current multispectral thermometries usually rely on an assumption model [4]–[6] and an experienced formula. If the formula cannot describe the object to be measured accurately, the error will exist in the measurement results of emissivity and the real temperature. It restricts the further development of multispectral thermometry.

This letter has established an activation- function-tunable neural network (AFT-CNNE) and a multistep searching method used to train the model in order to calculate object's continuous spectral emissivity. This model does not require the assumption of a fixed emissivity formula, and can realize on-line demarcation, and can also get very high precision for any object.

Manuscript received July 11, 2004 ; revised December 17, 2004.

C. Yang is with the Department of Electrical Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China, and also with the Department of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, P.R. China (e-mail: yangcl1@hit.edu.cn).

Y. Yu, D. Zhao, and G. Zhao are with the Department of Electrical Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China.

Digital Object Identifier 10.1109/TNN.2005.860837

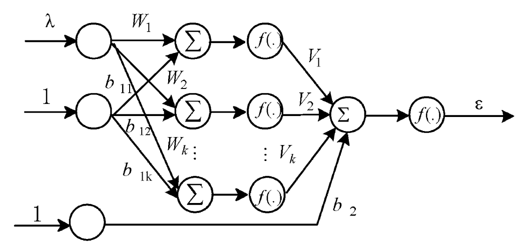

Fig. 1. Neural network structure.

II. CNNE Model Based on Fixed Activation Function

First of all, we need to use the neural network to establish the mapping relationship between emissivity and wavelength, that is $\varepsilon = \varepsilon(\lambda)$. As we cannot get the sample data of wavelength and emissivity from the actual measurement, so it is impossible to get the functional relationship model of ε directly from the BP network. Therefore, we combine the radiant temperature measurement theory with neural network to establish the combined neural network model. In this way, we can get the weight coefficients of the mapping relationship for $\varepsilon = \varepsilon(\lambda)$. Now we introduce the specific steps as follows.

A. Identification of the Nonlinear Relationship Between Wavelength and Emissivity

It has been proved in theory that a forward neural network with a hidden layer can approximate any nonlinear curve defined in a compact set in R^n with any precision. Since there is a nonlinear mapping relationship between the spectral emissivity and the wavelength of the object, a single hidden layer neural network is used to represent this relationship as shown in Fig. 1. The network has a structure with one input and one output, with k nodes. The input and output signals are denoted by λ and ε, respectively. The hidden layer errors are denoted by $b1_1, b1_2, \ldots, b1k$, respectively, and the output layer error is denoted by b_2. According to the theory of neural network, the activation function $f(.)$ of output and hidden layers can be either the same or different. Suppose the activation function is the same denoted by $f(.)$, then the relationship between the wavelength λ and the emissive ε can be obtained from Fig. 1 as follows.

$$\varepsilon = f\left(\left(\sum_{m=1}^{k} V_m f\left(W_m \lambda + b_{1m}\right)\right) + b_2\right). \tag{1}$$

Since the sample data of wavelength and emissivity cannot be obtained by using multispectral thermometry, the weights in above neural network cannot be calculated, which means the network has to be improved. Since the brightness temperature of each spectrum can be measured, a mathematical model can be constructed based on the brightness temperature so as to obtain the best approximation to the weights of the object's emissivity, and then the nonlinear mapping relationship between the emissivity and the wavelength can be achieved. The CNNE model below can realize above purpose.

B. Construction of the Combined Neural Network Model

Let the number of channels of the multispectral radiation thermometer be n, then, according to the theory of radiation thermometry, the relationship among the real object temperature T, the brightness