Capturing Correlations Among Facial Parts for Facial Expression Analysis

Caifeng Shan, Shaogang Gong, and Peter W. McOwan
Department of Computer Science
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
{cfshan, sgg, pmco}@dcs.qmul.ac.uk

Abstract

Capturing and analyzing the correlations among facial parts are important for interpreting facial behaviors precisely. In this paper, we exploit Canonical Correlation Analysis (CCA) to model the correlations of facial parts for facial expression analysis. We propose a Matrix-based Canonical Correlation Analysis (MCCA) for better correlation analysis on 2D image or matrix data in general. Extensive experiments have shown that compared to the traditional CCA, MCCA models more accurately correlations among image data with more compact representation using much fewer canonical factors.

1 Introduction

Automatic facial expression analysis has attracted much attention in recent years [13]. As facial muscles are contracted in unison to display expressions, different facial parts have strong correlations. Capturing and analyzing the correlations among facial parts are important for interpreting facial expressions precisely. Most of the existing work on facial expression analysis [4, 14, 2, 11] did not explicitly model the correlations between facial parts. In this paper, we employ Canonical Correlation Analysis (CCA) (Section 2), a statistical technique that is well suited for relating two sets of signals, to model correlations of facial parts for facial expression analysis.

CCA was developed [8] for measuring linear relationships between two vector variables. It finds pairs of base vectors (i.e., canonical factors) for two variables such that the correlations between the projections of the variables onto these canonical factors are mutually maximized. Recently CCA has been applied to computer vision problems [1, 12, 7, 10, 5]. Borga [1] adopted CCA to find corresponding points in stereo images. Melzer *et al.*[12] applied CCA to model the relation between an object's poses with raw brightness images for pose estimation. Like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), CCA also reduces the dimensionality of the original variables, since only a few factor pairs are normally needed to represent the relevant information. However, they serve different purposes: whilst PCA aims to minimize the reconstruction error and LDA derives a discriminant function that maximizes between-class scatter and minimize within-class scatter, CCA seeks directions for two sets of variables to maximize their correlations, so it is better suited for regression tasks. It has been shown that CCA outperforms PCA for regression tasks [12]. Recently Donner *et al.*[5] presented a fast Active Appearance Model search algorithm, which uses reduce-rank regression estimates obtained by CCA, instead of standard linear least-square regression estimates.

BMVC 2007 doi:10.5244/C.21.51

In the existing work, when applying CCA to image data, the original two-dimensional images have to be reshaped into one-dimensional vectors, as the traditional CCA is based on the vector-space model. However, this matrix-to-vector operation leads to two main problems. Firstly, the intrinsic 2D structure of image matrices is removed, so the spatial information stored therein is discarded. CCA based on these vectors can not fully capture correlations among the original 2D image data. Secondly, each image sample is modeled as a high-dimensional vector so that a large number of training samples are needed to yield a reliable estimation of the underlying data distribution. However, in reality, very limited number of training data are usually available. Actually these problems are shared by other subspace methods such as PCA and LDA. Recently some methods have been proposed to extend these vector-based methods for 2D matrices or high-order tensors [16, 17, 3, 15]. However, all these existing matrix-based methods were developed for learning in one set of variables, and not suited for measuring relationships between two set of variables.

To address these problems, we introduce a novel Matrix-based Canonical Correlation Analysis (MCCA) for better correlation analysis of 2D image or matrix data in general (Section 3). MCCA takes a 2D matrix based data representation rather than the 1D vector based representation in classical CCA. So the collection of data is represented as a set of matrices, instead of a single large matrix. MCCA seeks canonical factors in two dimensions to maximize the correlations between two sets of matrices. Unlike classical CCA, there is no closed-form solution for the optimization problem in MCCA. Instead, we propose an iterative solution with a convergence proof. We evaluate the proposed MCCA in capturing correlations of facial parts for facial expression analysis (Section 4). Experimental results demonstrate that MCCA can better measure correlations in 2D image data, providing superior performance in regression and recognition tasks, whilst requiring much fewer canonical factors. We notice that more recently Zou *et al.*[18] introduced a 2DCCA by simply replacing the image vector with image matrix in computing the variance matrices. Their approach is different to ours both in concept and algorithmic design; moreover, they addressed the correlations between image sets and their label matrices, instead of two sets of images.

2 Canonical Correlation Analysis

Given two zero-mean random variables $\mathbf{x} \in R^m$ and $\mathbf{y} \in R^n$, CCA finds pairs of directions $\mathbf{w}_x$ and $\mathbf{w}_y$ that maximize the correlation between the projections $x = \mathbf{w}_x^T \mathbf{x}$ and $y = \mathbf{w}_y^T \mathbf{y}$ (x and y are called *canonical variates*). More formally, CCA maximizes the function:

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\mathbf{w}_x^T \mathbf{x}\mathbf{y}^T \mathbf{w}_y]}{\sqrt{E[\mathbf{w}_x^T \mathbf{x}\mathbf{x}^T \mathbf{w}_x]E[\mathbf{w}_y^T \mathbf{y}\mathbf{y}^T \mathbf{w}_y]}} = \frac{\mathbf{w}_x^T \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_y^T \mathbf{C}_{yy} \mathbf{w}_y}} \quad (1)$$

where $\mathbf{C}_{xx} \in R^{m \times m}$ and $\mathbf{C}_{yy} \in R^{n \times n}$ are the *within-set covariance matrices* of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $\mathbf{C}_{xy} \in R^{m \times n}$ denotes their *between-sets covariance matrix*. A number of at most $k = \min(m,n)$ canonical factor pairs $\langle \mathbf{w}_x^i, \mathbf{w}_y^i \rangle, i = 1, \ldots, k$ can be obtained by successively solving $\arg\max_{\mathbf{w}_x^i, \mathbf{w}_y^i} \{\rho\}$ subject to $\rho(\mathbf{w}_x^j, \mathbf{w}_x^i) = \rho(\mathbf{w}_y^j, \mathbf{w}_y^i) = 0$ for $j = 1, \ldots, i-1$, i.e., the next pair of $\langle \mathbf{w}_x, \mathbf{w}_y \rangle$ are orthogonal to the previous ones.

The maximization problem can be solved by setting the derivatives of Eqn. (1), with

respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, equal to zero, resulting in the eigenvalue equations as:

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x = \rho^2\mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y = \rho^2\mathbf{w}_y \end{cases} \tag{2}$$

Matrix inversions need to be performed in Eqn. (2), leading to numerical instability if $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are rank deficient. Alternatively, $\mathbf{w}_x$ and $\mathbf{w}_y$ can be obtained by computing principal angles, as CCA is the statistical interpretation of principal angles between two linear subspace [6] (see [10] for details).

3 Matrix-based Canonical Correlation Analysis

We present an approach to perform canonical correlation analysis on 2-dimensional images or matrices in general. Given two matrix variables $\mathbf{A} \in R^{m\times n}$ and $\mathbf{B} \in R^{j\times k}$ (we assume the variables are both zero-mean), MCCA finds pairs of directions $\mathbf{v}_a \in R^m$, $\mathbf{w}_a \in R^n$, $\mathbf{v}_b \in R^j$ and $\mathbf{w}_b \in R^k$ that maximize the correlation between the projections $a = \mathbf{v}_a^T\mathbf{A}\mathbf{w}_a$ and $b = \mathbf{v}_b^T\mathbf{B}\mathbf{w}_b$. Mathematically, we can formulate this as the following maximization problem: find optimal $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ that maximize

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^T\mathbf{A}\mathbf{w}_a\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T\mathbf{A}\mathbf{w}_a\mathbf{w}_a^T\mathbf{A}^T\mathbf{v}_a]E[\mathbf{v}_b^T\mathbf{B}\mathbf{w}_b\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b]}} \tag{3}$$

Here $\mathbf{v}_a$ ($\mathbf{v}_b$) and $\mathbf{w}_a$ ($\mathbf{w}_b$) are canonical factors in two dimensions, acting as a two-sided linear transformation on the data in matrix form. To our knowledge, there is no closed-form solution for the maximization problem in Eqn. (3). A key observation, which leads to an iterative algorithm for the computation of $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$, is stated in the following Lemma:

Lemma 1 *Let $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ be the optimal solution to the maximization problem in Eqn. (3), then*
(1) Given $\mathbf{w}_a$ and $\mathbf{w}_b$, $\mathbf{v}_a$ and $\mathbf{v}_b$ can be obtained as canonical factors of two variables $\mathbf{a}' \in R^m$ and $\mathbf{b}' \in R^j$, where $\mathbf{a}' = \mathbf{A}\mathbf{w}_a$ and $\mathbf{b}' = \mathbf{B}\mathbf{w}_b$.
(2) Given $\mathbf{v}_a$ and $\mathbf{v}_b$, $\mathbf{w}_a$ and $\mathbf{w}_b$ can be obtained as canonical factors of two variables $\mathbf{a}'' \in R^n$ and $\mathbf{b}'' \in R^k$, where $\mathbf{a}'' = \mathbf{A}^T\mathbf{v}_a$ and $\mathbf{b}'' = \mathbf{B}^T\mathbf{v}_b$.

Proof (1) $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ maximize Eqn. (3), which can be rewritten as

$$\rho = \frac{E[\mathbf{v}_a^T\mathbf{a}'\mathbf{b}'^T\mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T\mathbf{a}'\mathbf{a}'^T\mathbf{v}_a]E[\mathbf{v}_b^T\mathbf{b}'\mathbf{b}'^T\mathbf{v}_b]}} \tag{4}$$

where $\mathbf{a}' = \mathbf{A}\mathbf{w}_a$ and $\mathbf{b}' = \mathbf{B}\mathbf{w}_b$. Hence, given $\mathbf{w}_a$ and $\mathbf{w}_b$, the maximum of Eqn. (4) is achieved by solving canonical correlation analysis on the variables $\mathbf{a}'$ and $\mathbf{b}'$ (by the definition of CCA in Eqn. (1)). So $\mathbf{v}_a$ and $\mathbf{v}_b$ can be obtained as canonical factors of $\mathbf{a}'$ and $\mathbf{b}'$.

(2) Similarly, Eqn. (3) can also be rewritten as

$$\rho = \frac{E[\mathbf{w}_a^T\mathbf{a}''\mathbf{b}''^T\mathbf{w}_b]}{\sqrt{E[\mathbf{w}_a^T\mathbf{a}''\mathbf{a}''^T\mathbf{w}_a]E[\mathbf{w}_b^T\mathbf{b}''\mathbf{b}''^T\mathbf{w}_b]}} \tag{5}$$

where $\mathbf{a}'' = \mathbf{A}^T\mathbf{v}_a$ and $\mathbf{b}'' = \mathbf{B}^T\mathbf{v}_b$. Hence, given $\mathbf{v}_a$ and $\mathbf{v}_b$, the maximum of Eqn. (5) is achieved by solving canonical correlation analysis on the variables $\mathbf{a}''$ and $\mathbf{b}''$. So $\mathbf{w}_a$ and $\mathbf{w}_b$ can be obtained as canonical factors of $\mathbf{a}''$ and $\mathbf{b}''$. This completes the proof of the lemma.

By the above Lemma, we present an iterative procedure for computing $\mathbf{v}_a$, $\mathbf{w}_a$, $\mathbf{v}_b$ and $\mathbf{w}_b$ as follows: given the initial choice of $\mathbf{w}_a$ and $\mathbf{w}_b$, we can compute $\mathbf{v}_a$ and $\mathbf{v}_b$ by computing canonical factors of $\mathbf{a}'$ and $\mathbf{b}'$; with the computed $\mathbf{v}_a$ and $\mathbf{v}_b$ (corresponding to the largest canonical correlation), we can then compute $\mathbf{w}_a$ and $\mathbf{w}_b$ by computing canonical factors of $\mathbf{a}''$ and $\mathbf{b}''$, and $\mathbf{w}_a$ and $\mathbf{w}_b$ (corresponding to the largest canonical correlation) will be used in next iteration. The procedure can be repeated until convergence. In this way, a number of at most $q = \min(m, j)$ left-side canonical factor pairs $\langle \mathbf{v}_a^1, \mathbf{v}_b^1 \rangle, \ldots, \langle \mathbf{v}_a^q, \mathbf{v}_b^q \rangle$ and a number of at most $p = \min(n, k)$ right-side canonical factor pairs $\langle \mathbf{w}_a^1, \mathbf{w}_b^1 \rangle, \ldots, \langle \mathbf{w}_a^p, \mathbf{w}_b^p \rangle$ can be obtained. The pseudo-code of the above iterative procedure is given in Algorithm 1, where `CCA`$(\mathbf{a}, \mathbf{b})$ computes the canonical factors and canonical correlations of the variables $\mathbf{a}$ and $\mathbf{b}$.

Algorithm 1: MCCA

1 Obtain initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$ for $\mathbf{w}_a$ and $\mathbf{w}_b$, and set $\rho^{(0)} \leftarrow -1$ and $i \leftarrow 0$;
2 **repeat**
3 $\quad i \leftarrow i + 1$;
4 $\quad (\mathbf{v}_a^s, \mathbf{v}_b^s, \rho^s) \leftarrow$ `CCA` $(\mathbf{A} * \mathbf{w}_a^{(i-1)}, \mathbf{B} * \mathbf{w}_b^{(i-1)})$;
$\quad$ /* $s = 1, \ldots, q$ */
5 $\quad \mathbf{v}_a^{(i)} \leftarrow \mathbf{v}_a^1$, $\mathbf{v}_b^{(i)} \leftarrow \mathbf{v}_b^1$, $\rho^{(i)} \leftarrow \rho^1$;
6 $\quad (\mathbf{w}_a^t, \mathbf{w}_b^t, \rho^t) \leftarrow$ `CCA` $(\mathbf{A}^T * \mathbf{v}_a^{(i)}, \mathbf{B}^T * \mathbf{v}_b^{(i)})$;
$\quad$ /* $t = 1, \ldots, p$ */
7 $\quad \mathbf{w}_a^{(i)} \leftarrow \mathbf{w}_a^1$, $\mathbf{w}_b^{(i)} \leftarrow \mathbf{w}_b^1$, $\rho^{(i)} \leftarrow \rho^1$;
8 **until** $\rho^{(i)} - \rho^{(i-1)} < \varepsilon$;
9 $\mathbf{V}_a \leftarrow [\mathbf{v}_a^1, \ldots, \mathbf{v}_a^q]$, $\mathbf{V}_b \leftarrow [\mathbf{v}_b^1, \ldots, \mathbf{v}_b^q]$;
10 $\mathbf{W}_a \leftarrow [\mathbf{w}_a^1, \ldots, \mathbf{w}_a^p]$, $\mathbf{W}_b \leftarrow [\mathbf{w}_b^1, \ldots, \mathbf{w}_b^p]$;

3.1 Proof of Convergence

The convergence of MCCA follows, since correlation coefficient ρ is bounded between -1 and 1 from its definition, as stated in the following theorem:

Theorem 2 *The MCCA algorithm monotonically non-decreases the value of correlation coefficient* ρ*, hence it converges in the limit.*

Proof Given $\mathbf{w}_a^{(i-1)}$, $\mathbf{w}_b^{(i-1)}$ and $\rho^{(i-1)}$ obtained in Line 7, `CCA` $(\mathbf{A} * \mathbf{w}_a^{(i-1)}, \mathbf{B} * \mathbf{w}_b^{(i-1)})$ in Line 4 finds optimal $\mathbf{v}_a$ and $\mathbf{v}_b$ that maximize

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^T \mathbf{A}\mathbf{w}_a^{(i-1)} {\mathbf{w}_b^{(i-1)}}^T \mathbf{B}^T \mathbf{v}_b]}{\sqrt{E[\mathbf{v}_a^T \mathbf{A}\mathbf{w}_a^{(i-1)} {\mathbf{w}_a^{(i-1)}}^T \mathbf{A}^T \mathbf{v}_a] E[\mathbf{v}_b^T \mathbf{B}\mathbf{w}_b^{(i-1)} {\mathbf{w}_b^{(i-1)}}^T \mathbf{B}^T \mathbf{v}_b]}} \tag{6}$$

Apparently, the value of $\rho^{(i-1)}$ is derived as

$$\begin{aligned}\rho^{(i-1)} &= \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} \\ &= \frac{E[\mathbf{v}_a^{(i-1)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i-1)}]}{\sqrt{E[\mathbf{v}_a^{(i-1)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_a^{(i-1)^T}\mathbf{A}^T\mathbf{v}_a^{(i-1)}]E[\mathbf{v}_b^{(i-1)^T}\mathbf{B}\mathbf{w}_b^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i-1)}]}}\end{aligned} \tag{7}$$

which is less or equal to the maximized canonical correlation that $\texttt{CCA}(\mathbf{A}*\mathbf{w}_a^{(i-1)}, \mathbf{B}*\mathbf{w}_b^{(i-1)})$ finds. so the derived $\rho^{(i)}$ in Line 5 is no less than $\rho^{(i-1)}$. With regard to the first iteration, given any initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$, the canonical correlation $\rho^{(1)}$ derived by $\texttt{CCA}(\mathbf{A}*\mathbf{w}_a^{(0)}, \mathbf{B}*\mathbf{w}_b^{(0)})$ is no less than -1 ($\rho^{(0)}$). Therefore, the update of ρ in Line 5 do not decrease its value, since the computed ρ is locally optimal.

Similarly, given $\mathbf{v}_a^{(i)}$, $\mathbf{v}_b^{(i)}$ and $\rho^{(i)}$ obtained in Line 5, $\texttt{CCA}(\mathbf{A}^T*\mathbf{v}_a^{(i)}, \mathbf{B}^T*\mathbf{v}_b^{(i)})$ in Line 6 finds optimal $\mathbf{w}_a$ and $\mathbf{w}_b$ that maximize

$$\rho = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b^{(i)}]}{\sqrt{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a\mathbf{w}_a^T\mathbf{A}^T\mathbf{v}_a^{(i)}]E[\mathbf{v}_b^{(i)^T}\mathbf{B}\mathbf{w}_b\mathbf{w}_b^T\mathbf{B}^T\mathbf{v}_b^{(i)}]}} \tag{8}$$

Apparently, the value of $\rho^{(i)}$ in Line 5 is derived as

$$\rho^{(i)} = \frac{E[ab]}{\sqrt{E[a^2]E[b^2]}} = \frac{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i)}]}{\sqrt{E[\mathbf{v}_a^{(i)^T}\mathbf{A}\mathbf{w}_a^{(i-1)}\mathbf{w}_a^{(i-1)^T}\mathbf{A}^T\mathbf{v}_a^{(i)}]E[\mathbf{v}_b^{(i)^T}\mathbf{B}\mathbf{w}_b^{(i-1)}\mathbf{w}_b^{(i-1)^T}\mathbf{B}^T\mathbf{v}_b^{(i)}]}} \tag{9}$$

which is less or equal to the maximized canonical correlation that $\texttt{CCA}(\mathbf{A}^T*\mathbf{v}_a^{(i)}, \mathbf{B}^T*\mathbf{v}_b^{(i)})$ finds. So the update of ρ in Line 7 do not decrease its value too. Therefore, the MCCA optimization process monotonically non-decreases the ρ value, and converges in the limit. This completes the proof of the theorem.

The convergence of the MCCA algorithm was also confirmed experimentally. We show some examples of iterative learning in Fig. 1 and Fig. 2, where each example is for the learning on a different training set. We can observe that the value of ρ becomes stable after at most 20-30 iterations. We also found that any variation on the initial choice of $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$ has almost no effect on convergence (as observed in Fig. 2). The fast and stable convergence keeps the training cost low.

3.2 Effect of the Initial Choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$.

Theoretically, our solution to MCCA is only locally optimal. This solution depends on the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$. However, in practice this does not have any ill-effect. We conducted extensive experiments using different choices for $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$, and found that, for image datasets, MCCA always converges to a similar (if not identical) solution regardless of the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$.

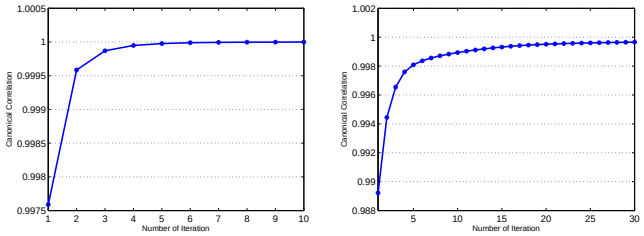

Figure 1: Convergence property of MCCA.

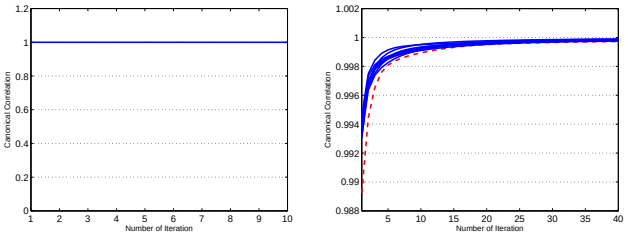

Figure 2: Sensitivity of MCCA to the initial choice $\mathbf{w}_a^0$ and $\mathbf{w}_b^0$: the ten solid curves correspond to the ten runs with random initializations, and the dash curve corresponds to $\mathbf{w}_a^0 = \mathbf{w}_b^0 = (1,0,\ldots,0)^T$ (the dash curve in the *left* side is identical with solid curves, so is not visible).

We show two typical results in Fig. 2, where the horizontal axis is the number of iterations and the vertical axis is the value of ρ. Each sub-figure is the results on a different training set. We run MCCA with 10 randomly generated $\mathbf{w}_a^{(0)}$'s and $\mathbf{w}_b^{(0)}$'s, and another initialization $\mathbf{w}_a^{(0)} = \mathbf{w}_b^{(0)} = (1,0,\ldots,0)^T$. For the left side of Fig. 2, we can observe that MCCA converges within two iterations for all eleven initial choices with the specified threshold ($\varepsilon = 10^{-5}$), and also converges to the same solution. In the right side of Fig. 2, MCCA converges slower. For all different initial choices, MCCA converges within 20-30 iterations with the threshold $\varepsilon = 10^{-5}$, and converges to very similar solutions. The difference between the values of final ρ is very small ($< 1.6 \times 10^{-4}$). These experiments demonstrate that, for image datasets, MCCA always converges to a similar (if not identical) solution regardless of the initial choice $\mathbf{w}_a^{(0)}$ and $\mathbf{w}_b^{(0)}$. We used the initial choice $\mathbf{w}_a^{(0)} = \mathbf{w}_b^{(0)} = (1,0,\ldots,0)^T$ in our experiments.

4 Experiments

As a case study, we investigate correlations between the mouth part (Mouth) and the right eye part (Eye) (as shown in Fig. 3). These two parts have strong and a range of correlations corresponding to facial expressions. We conducted experiments on the Cohn-Kanade database [9] and face expression image sequences we captured. We manually normalized the faces based on three feature points, centers of the two eyes and the mouth, using affine transformation. In the normalized facial images (110×150 pixels), the mouth part is 53×68 pixels, and the eye part is 45×51 pixels.

Figure 3: A case study on correlations between the mouth and the right eye facial parts.

4.1 Facial Parts Synthesis

We wish to reconstruct (synthesize) Mouth from Eye or vice versa using MCCA based regression. Specifically, to reconstruct image **B** from image **A**, we first employ MCCA to establish their relationship, finding optimal projection directions in the sense of correlation, and then map A to the leading canonical variates by discarding directions with low canonical correlation. Finally we perform regression of **B** by taking these leading canonical variates of **A**. The procedure of synthesis is as follows.

1. Compute the leading factor pairs $\mathbf{V}_a, \mathbf{W}_a, \mathbf{V}_b, \mathbf{W}_b$ from N pairs of samples $\mathscr{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N\}$ and $\mathscr{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_N\}$.

2. Map $\mathbf{A}_i$ $(i = 1, \ldots, N)$ to the reduced correlation space $\widetilde{\mathbf{A}}_i = \mathbf{V}_a^T \mathbf{A}_i \mathbf{W}_a$.

3. Reshape 2D matrices $\widetilde{\mathbf{A}}_i$ and $\mathbf{B}_i$ to 1D vectors $\tilde{\mathbf{a}}_i$ and $\mathbf{b}_i$, and form data matrices $\widetilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_N]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]$; then compute the regression matrix $\mathbf{R} = (\widetilde{\mathbf{A}}^T)^{-1}\mathbf{B}^T$.

4. Given a new input $\mathbf{A}_{new}$, the corresponding $\mathbf{B}_{new}$ is reconstructed by:

$$\widetilde{\mathbf{A}}_{new} = \mathbf{V}_a^T \mathbf{A}_{new} \mathbf{W}_a, \quad \widetilde{\mathbf{A}}_{new} \rightarrow \tilde{\mathbf{a}}_{new} \tag{10}$$

$$\mathbf{b}_{new} = \mathbf{R}^T \tilde{\mathbf{a}}_{new}, \quad \mathbf{b}_{new} \rightarrow \mathbf{B}_{new} \tag{11}$$

Here $\widetilde{\mathbf{A}}_{new} \rightarrow \tilde{\mathbf{a}}_{new}$ represents reshaping 2D matrix $\widetilde{\mathbf{A}}_{new}$ to 1D vector $\tilde{\mathbf{a}}_{new}$, and $\mathbf{b}_{new} \rightarrow \mathbf{B}_{new}$ is reshaping 1D vector $\mathbf{b}_{new}$ to 2D matrix $\mathbf{B}_{new}$. This reconstruction procedure is not limited to facial parts but can also be generally applied to other types of image synthesis.

We selected more than 10 subjects from the Cohn-Kanade database, each of which has around 70$\sim$140 images of different facial expressions, in addition to the image sequences we captured. For the image set of each subject, we randomly sampled one tenth of the images as the testing set, and the remaining images as the training set. We applied MCCA, CCA, and the standard linear least-squares regression (SR) approach to synthesize Mouth from Eye and vice versa on the testing set. We used 10 randomly selected training/testing combinations for reporting reconstruction errors. We observe that MCCA performs better than CCA and SR in reconstructing one facial part from another. Moreover, MCCA requires much fewer canonical factors to obtain better reconstruction results. We report the reconstruction results for six random selected subjects in Table 1, where the optimal average pixel errors (with standard deviation) and the corresponding dimensions of canonical factors used are reported. To clearly compare the three methods, we plot bar graphs of average pixel errors and the dimensions of the canonical factors used in MCCA/CCA in Fig. 4. Some reconstruction examples are shown in Fig. 5 (A supplementary video demonstration is available at *http://www.dcs.qmul.ac.uk/$\sim$cfshan/research/cca.html*).

Subject	Algorithm	Eye → Mouth		Mouth → Eye	
		Pixel Errors	Dims	Pixel Errors	Dims
(1)	2DCCA	11.2±2.0	11*6	8.8±1.2	2*23
	CCA	16.7±4.4	139	13.1±4.0	139
	SR	17.3±3.5	-	14.7±4.8	-
(2)	2DCCA	8.5±2.5	9*6	8.4±1.8	5*10
	CCA	13.0±6.0	119	10.7±3.6	119
	SR	12.4±5.3	-	10.7±3.2	-
(3)	2DCCA	13.4±5.5	15*3	10.0±3.3	28*1
	CCA	16.2±8.8	96	11.6±5.9	96
	SR	16.1±8.8	-	12.0±6.5	-
(4)	2DCCA	16.3±5.0	39*1	19.5±6.0	17*3
	CCA	24.5±9.8	96	25.4±18.3	96
	SR	23.7±7.6	-	26.1±18.8	-
(5)	2DCCA	9.9±1.8	14*2	10.5±2.5	22*2
	CCA	12.9±4.4	85	11.0±3.1	85
	SR	14.2±4.3	-	11.0±2.9	-
(6)	2DCCA	13.8±2.4	28*1	12.6±2.9	18*2
	CCA	17.2±8.5	77	15.7±6.2	77
	SR	15.1±4.9	-	13.9±6.1	-

Table 1: The reconstruction results for six subjects: the optimal average pixel errors (with standard deviation) of the three algorithms, and the corresponding dimensions of canonical factors used in MCCA and CCA.

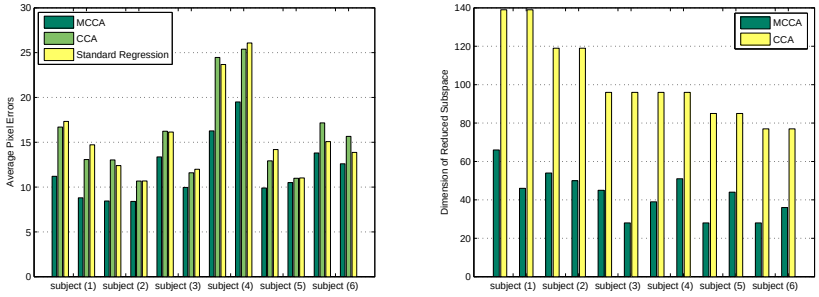

Figure 4: (*left*) Reconstruction errors of the three algorithms; (*right*) Dimensions of canonical factors used in MCCA and CCA. (Two groups bars for each subject: the left is 'Eye → Mouth' and the right is 'Mouth → Eye'.)

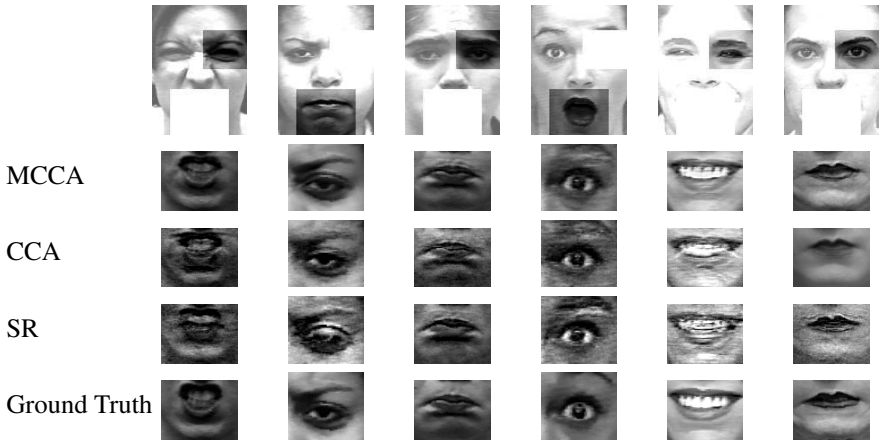

Figure 5: Some examples of facial parts synthesis using MCCA, CCA, and SR.

It is compelling that MCCA outperforms CCA and SR consistently in facial parts synthesis. Crucially, as observed in Fig. 4, the dimension of canonical factors needed in MCCA is always less than 50% of that of CCA. So MCCA can describe correlations among facial parts with better accuracy using much less canonical factors. The superior performance of MCCA credits to its ability to preserve the intrinsic 2D spatial structure and capture the correlation store therein, and its robustness with limited number of training data. The strength of MCCA is also reflected by the average standard deviation. As shown in Table 1, MCCA always produces the smallest deviation, which suggests that MCCA is much more robust. Compared to SR, where the full-rank regression matrix has to be estimated from a limited number of noisy training images, the MCCA based reduced-rank regression provides more reliable parameter estimates by taking advantage of correlations between the image sets, leading to better accuracy and robustness.

4.2 Facial Expression Recognition

We also conducted facial expression recognition experiments based on correlations between Mouth and Eye. The basic idea is that these two parts have distinctive correlations for different expressions, so the correlations modeled by MCCA should provide discriminant information for expression classification. Given image sets of different facial expressions $I_1, \ldots, I_c$ (c is the number of classes), we derive the leading factor pairs $(\mathbf{V}_a^i, \mathbf{W}_a^i, \mathbf{V}_b^i, \mathbf{W}_b^i), i = 1 \ldots c$ of parts Mouth (denoted by $\mathbf{B}$) and Eye (denoted by $\mathbf{A}$) for each class using MCCA. We then compute the regression parameters for reconstructing $\mathbf{B}$ from $\mathbf{A}$ in the reduced correlation space in the training set. Given a test image $\mathbf{I}_{new}$ of an unknown class, we map its Eye $\mathbf{A}_{new}$ and Mouth $\mathbf{B}_{new}$ to the reduced correlation space of class i as $\widetilde{\mathbf{A}}_i = (\mathbf{V}_a^i)^T \mathbf{A}_{new} \mathbf{W}_a^i$ and $\widetilde{\mathbf{B}}_i = (\mathbf{V}_b^i)^T \mathbf{B}_{new} \mathbf{W}_b^i$, and then calculate the error $err(i)$ of reconstructing $\widetilde{\mathbf{B}}_i$ from $\widetilde{\mathbf{A}}_i$ with the regression parameters of this correlation space. After computing the reconstruction error of each class $err(i), i = 1 \ldots c$, we classify the test image as the class having the smallest reconstruction error

$$\hat{i} = \arg\min_i err(i) \tag{12}$$

For our experiments, we selected 732 image of basic emotions (Anger, Disgust, Joy, and Surprise) from the Cohn-Kanade database. The sequences come from 96 subjects, with 1 to 4 emotions per subject. We first considered a 2-class (Joy and Surprise) recognition problem, then included Anger for a 3-class problem, and finally considered four expressions for classification (incrementally making the recognition task harder). To evaluate generalization performance, a 10-fold Cross-Validation testing scheme was adopted. The recognition results using MCCA and CCA are reported in Table 2. We can observe that expressions can be better classified using MCCA, demonstrating again that MCCA outperform CCA in capturing correlations in facial parts. It is also evident that by modeling correlations between only two facial parts, the recognition accuracy degrades quickly for multi-class recognition. By considering correlations of multiple facial parts, we should be able to improve these recognition results.

	2-Class	3-Class	4-Class
MCCA	96.1±3.6	80.8±6.4	67.9±4.8
CCA	63.2±10.5	55.6±7.8	48.7±6.7

Table 2: Facial expression recognition based on correlations of Mouth and Eye modeled by MCCA and CCA.

5 Conclusions

Experimental results have shown that the proposed MCCA can better model correlations among image data with much fewer canonical factors. The underlying reason is that MCCA is able to preserve and utilize the intrinsic 2D spatial structure in image data. MCCA is still a linear technique, however, so it cannot effectively deal with higher-order statistics among image data. Our future work will focus on formulating nonlinear MCCA.

References

[1] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linkoping University, SE-581 83 Linkoping, Sweden, 1998. Dissertation No 531.

[2] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *CVIU*, 91:160–187, 2003.

[3] G. Dai and D.-Y. Yeung. Tensor embedding methods. In *National Conference on Artificial Intelligence (AAAI)*, pages 330–335, 2006.

[4] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE PAMI*, 21(10):974–989, October 1999.

[5] R. Donner, M. Reiter, G. Langs, P. Peloscheck, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE PAMI*, 28(10):1690–1694, October 2006.

[6] G. H. Golub and H. Zha. The canonical correlations of matrix pairs and their numerical computation. Technical report, 1992.

[7] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 8:321–377, 1936.

[9] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE FG*, 2000.

[10] T-K. Kim, J. Kittler, and R. Cipolla. Learning discriminative canonical correlations for object recognition with image sets. In *ECCV*, pages 251–262, 2006.

[11] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, June 2006.

[12] T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 39(9):1961–1973, 2003.

[13] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: the state of art. *IEEE PAMI*, 22(12):1424–1445, 2000.

[14] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE PAMI*, 23(2):97–115, February 2001.

[15] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang. Multilinear discriminant analysis for face recognition. *IEEE Transactions on Image Processing*, 16(1):212–220, January 2007.

[16] J. Yang, A. F. Zhang, D. Frangi, and J. Yang. Two dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE PAMI*, 26(1):131–137, 2004.

[17] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191, November 2005.

[18] C. Zou, N. Sun, Z. Ji, and L. Zhao. 2dcca: A novel method for small sample size face recognition. In *IEEE Workshop on Application of Computer Vision*, 2007.