



An Extension 2DPCA based Visual Feature Extraction Method for Audio-Visual Speech Recognition

Guanyong Wu, Jie Zhu

Department of Electronics Engineering, Shanghai Jiaotong University, Shanghai, China

{wuguanrong, zhujie}@sjtu.edu.cn

Abstract

Two dimensional principal component analysis (2DPCA) has been proposed for face recognition as an alternative to traditional PCA transform [1]. In this paper, we extend this approach to the visual feature extraction for audio-visual speech recognition (AVSR). First, a two-stage 2DPCA transform is conducted to extract the visual features. Then, the visemic linear discriminant analysis (LDA) is applied for post extraction processing. We investigate the presented method comparing with traditional PCA and 2DPCA. Experimental results show that the extension 2DPCA can reduce the dimension of 2DPCA and represent the testing mouth images better than PCA does; Moreover, 2DPCA+LDA needs less computation and has a better performance than PCA+LDA in the visual-only speech recognition; Finally, further experimental results demonstrate that our AVSR system using the extension 2DPCA method provides significant enhancement of robustness in noisy environments compared to the audio-only speech recognition.

Index Terms: two dimensional principal component analysis (2DPCA), Audio-visual speech recognition (AVSR), linear discriminant analysis (LDA), feature extraction

1. Introduction

In recent years, there has been much research on the technology of the audio-visual speech recognition (AVSR), which incorporates the visual information by using the video data of the speaker's mouth to improve the performance over audio-only speech recognition. During the last few decades, various visual feature extraction methods have been presented, which can be categorized into three classes [2], [3]: appearance (video pixel) based features; shape (lip contour) based features and the combination of both above. However, the dimension of the mouth images (region of interest, ROI) is generally too large to be quickly processed. In this correspondence, the first step of extracting visual features from ROI is the dimension reduction. The most popular methods achieve such reduction by using traditional image transforms: principal component analysis (PCA) [4-6], discrete cosine transform (DCT) [7-8], etc.

PCA, also known as Karhunen-Loeve transform (KLT), has been widely used for feature extraction in face recognition and AVSR. However, PCA needs to transform original ROI matrices into the same dimensional vectors, and then rely on these vectors to evaluate the covariance matrix and to determine the projector. Yang [1], [9] developed the 2DPCA transform for face recognition. Unlike PCA, in which the analysis and operation are based on one dimensional vector representation, 2DPCA directly computes the image covariance (scatter) matrix with 2D image matrix. Experimental results demonstrated that 2DPCA was superior to the traditional PCA [1], [10].

Despite of the successful applications of 2DPCA in face recognition, 2DPCA needs more coefficients to represent a ROI image than PCA does, and 2DPCA can not reduce the high dimensionality. Xu [11] studied the problem and proposed a two-stage 2DPCA, named as parallel image matrix compression (PIMC) method to reduce the redundancy among columns and rows. Similarly, Zuo [12] proposed a similar method named as bidirectional PCA (BDPCA), which also used the 2DPCA twice by left multiplying and right multiplying two projection matrices simultaneously. In fact, Both of PIMC and BDPCA are the extensions to the 2DPCA method.

In this paper, motivated by the above studies, we investigate the extension 2DPCA transform to ROI images for AVSR system. Then, visemic based LDA is applied to the post extraction processing. Although 2DPCA has been used in face recognition domain, to the best of the author's knowledge, this is the first application of the extension 2DPCA method in the AVSR domain.

This paper is organized as follows: section 2 describes the PCA and 2DPCA transforms; section 3 presents the visual feature extraction method: the extension 2DPCA plus LDA; Section 4 gives the experimental results and a summary is in section 5.

2. PCA and 2DPCA Transforms

2.1. Principal Component Analysis

In a PCA transform, an image matrix of size $m \times n$ must be transformed into a one-dimensional vector X ($D = m \times n$) in advance. The transform can be rewritten as $Y = PX$, where P is the projection matrix. Given a set of M training images, $\{X^1, X^2, \dots, X^M\}$, the covariance matrix of PCA is defined by

$$S_c = \frac{1}{M} \sum_{j=1}^M (X^j - \mu)(X^j - \mu)^T = \frac{1}{M} \sum_{j=1}^M X^j \cdot X^{jT} - \mu\mu^T \quad (1)$$

Where $\mu = \frac{1}{M} \sum_{j=1}^M X^j$ denotes the mean vector. It is easy to

verify that S_c is a $D \times D$ nonnegative definite matrix. We choose the first d largest eigenvalues and the corresponding eigenvectors be $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_d\}$ and $V = \{v_1, v_2, \dots, v_d\}^T$. After the projection of the one-dimensional vector X onto the d -dimensional subspace V , we get the PCA-transformed vector, $Y = \{y_1, y_2, \dots, y_d\}^T$, which can be calculated by

$$Y = V(X - \mu) \quad (2)$$

So, it is easy to see that traditional PCA can reduce the dimension of an ROI image X from $D \times 1$ to $d \times 1$. We can reconstruct the original image by

$$\tilde{X} = V^T Y + \mu \quad (3)$$

2.2. 2DPCA Transform

Different from traditional PCA, for an $m \times n$ image matrix X , let U denote an n -dimensional unitary column vector. The 2DPCA transform is to project image X onto U by the linear transformation $Y = XU$. Thus, we get an m -dimensional projected vector Y . The covariance (scatter) matrix S is defined as [1], [9]

$$S = \frac{1}{M} \sum_{i=1}^M (X_i - \mu)(X_i - \mu)^T = \frac{1}{M} \sum_{i=1}^M X_i^T X_i - \mu^T \mu \quad (4)$$

Where $\mu = \frac{1}{M} \sum_{i=1}^M X_i$ is the mean matrix of all the training ROI images. We can also verify that S is an $n \times n$ nonnegative definite matrix. The optimal projection axes, U_1, U_2, \dots, U_d , actually are the orthonormal eigenvectors of S corresponding to the first d largest eigenvalues $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_d\}$.

$$Y_i = XU_i, i = 1, 2, \dots, d. \quad (5)$$

Where the projected vectors, Y_1, Y_2, \dots, Y_d are called the principal components of image X [1]. We can reconstruct the original image matrix by

$$\tilde{X} = YU^T = \sum_{i=1}^d Y_i U_i^T \quad (6)$$

3. Visual Feature Extraction

The visual feature extraction method we proposed in our audio-visual speech recognition system is shown in Fig1. Before the extraction of visual feature, the gray ROI images are down sampled to a normalized 32×32 rectangle. First, the extension 2DPCA is used to reduce the dimension of ROI images, then, the visual features (29.97Hz) were interpolated to make them occur at the same frame rate as the audio features (100Hz). Finally, the visemic LDA method is adopted to obtain the final visual feature vectors.

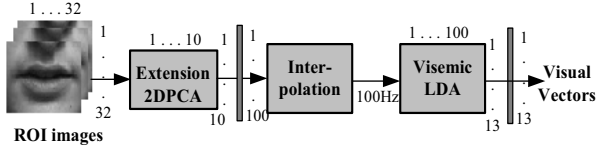


Figure 1: the extension 2DPCA based visual feature extraction for AVSR.

3.1. Extension of 2DPCA

From equations (2) and (5), it is clear that dimension of the 2DPCA feature vector ($m \times d$) is always much higher than PCA ($d \times 1$) when $m \gg d$. Here, we reduce the dimension by using a two-stage 2DPCA transform [11], [12], i.e. we left and right multiply two projection matrices simultaneously.

Given a training set of ROI images $\{X^1, X^2, \dots, X^M\}$ with a dimension $m \times n$, 2DPCA is first conducted to compute the projection directions $U = \{U_1, U_2, \dots, U_{d_1}\}$ by

$$W_i^j = X^j U_i, i = 1, 2, \dots, d_1, j = 1, 2, \dots, M. \quad (7)$$

$$S_1 = \frac{1}{M} \sum_{j=1}^M (X^j - \mu)^T (X^j - \mu) = \frac{1}{M} \sum_{j=1}^M X^{jT} X^j - \mu^T \mu \quad (8)$$

Where $U = \{U_1, U_2, \dots, U_{d_1}\}$ are the orthonormal eigenvectors of S_1 corresponding to the first d_1 largest eigenvalues. Then, for the j th projected feature matrix $W^j = \{W_1^j, W_2^j, \dots, W_{d_1}^j\}$, we

use the 2DPCA again to compute the projection direction $V = \{V_1, V_2, \dots, V_{d_2}\}$, and get $Y^j = \{Y_1^j, Y_2^j, \dots, Y_{d_2}^j\}$.

$$Y_i^j = V_k^T W_k^j, i = 1, 2, \dots, d_2, k = 1, 2, \dots, d_1. \quad (9)$$

$$S_2 = \frac{1}{M} \sum_{j=1}^M (W^j - \mu)(W^j - \mu)^T = \frac{1}{M} \sum_{j=1}^M W^j W^{jT} - \mu \mu^T \quad (10)$$

Where $V = \{V_1, V_2, \dots, V_{d_2}\}$ are the orthonormal eigenvectors of S_2 corresponding to the first d_2 largest eigenvalues.

With equations (7) and (9), the final projection matrix $Y^j = \{Y_1^j, Y_2^j, \dots, Y_{d_2}^j\}$ of ROI image X^j can be rewritten as

$$Y^j = V^T X^j U, j = 1, 2, \dots, M. \quad (11)$$

Specifically, similar to equation (2), we modify the equation (11) by subtracting each image from the mean image matrix.

$$Y^j = V^T (X^j - \mu) U, j = 1, 2, \dots, M. \quad (12)$$

Corresponding to equation (5), we named the matrix Y^j as the principal component (matrix) of image X^j . Hence, we can see that the extension 2DPCA reduces the dimension of X^j from $m \times n$ to $d_2 \times d_1$ (while 2DPCA reduces to $m \times d$).

Meanwhile, we also can reconstruct the original image by

$$\tilde{X}^j = V Y^j U^T + \mu \quad (13)$$

To evaluate the representation performance of the extension 2DPCA, we compare the reconstruction quality of PCA and the extension 2DPCA in Fig2. We first extract a training set from the video database containing 5000 mouth ROI images, and then reconstruct test images which are not included in the training set by the first d components. For the extension 2DPCA, we set $d_1 = d_2$. Fig2 (a) and (b) are two original testing ROI images, (c) and (e) are the reconstructed images by PCA, (d) and (f) are the reconstructed images by the extension 2DPCA. Fig2 intuitively shows the reconstructed images become clearer when the number of components is increased. When $d > 10$, we can see that the reconstructed images of (d) and (f) are more explicit than that of (c) and (e). With fewer components the extension 2DPCA can reconstruct the testing images in a satisfying quality.

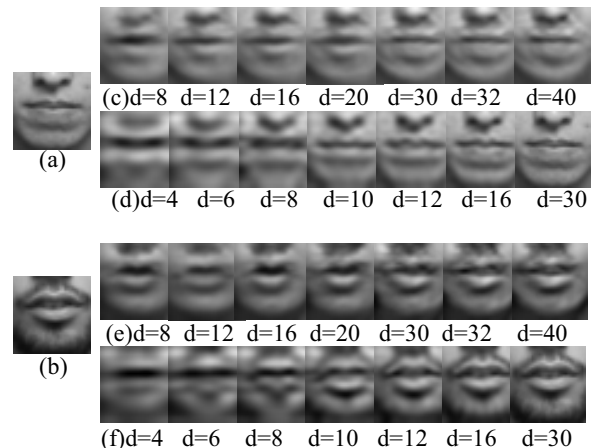


Figure 2: comparisons of the representation performance for testing ROI images of PCA and extension 2DPCA. (a) and (b) are original ROI images, (c) and (e) are reconstructed by PCA, (d) and (f) are reconstructed by the extension 2DPCA.

To further verify the reconstruction quality, we plot the Euclidean distance error curves between an original testing ROI image and the reconstructed images in Fig3. The curves

come down with the increasing use of components which reflects the changes corresponding to Fig2. From Fig3 we can see it clearly that the error value of extension 2DPCA decreases more quickly than that of PCA does, especially after the reconstruction using more than 8 components. When we use 32 components, the error of the extension 2DPCA is zero, which means there is no information lost during the reconstruction. Meanwhile, the error value of PCA is still relatively big. This shows the extension 2DPCA has a better representation performance of testing ROI images than PCA. Moreover, PCA transform is more time-consuming and more complex computational. Consequently we use 10 components ($d_1 = d_2 = 10$) in our experiments to extract visual features.

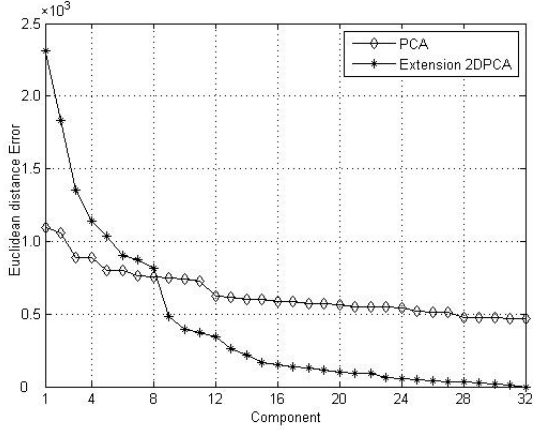


Figure3: Euclidean distance error using different components for reconstruction. X-axis denotes the components used, and Y-axis denotes the Euclidean distance error between the original testing ROI image and the reconstructed images using first d components.

3.2. Visemic Linear Discriminant Analysis

In the visual domain, the viseme is the basic unit of mouth movements that corresponds to the phoneme-the basic unit of speech in the acoustic domain. There are many acoustic sounds that are visually ambiguous in AVSR domain. These sounds are grouped into the same class that represents a viseme [13]. A reasonable class number mapping from the common 44 English phonemes to visemes is 13 [3]. In Fig1, we finally project the feature vectors on a 13 class visemic linear discriminant space to obtain the final visual feature vectors.

Here, the LDA is used to find the projection matrix $P = [P_1, P_2, \dots, P_m]^T$ that maximizes the ratio of between-class scatter S_B against within-class scatter S_W (Fisher's criterion).

$$\tilde{P} = \arg \max_P \frac{|PS_B P^T|}{|PS_W P^T|} \quad (14)$$

Given a set of $c = 1, 2, \dots, C$ classes, after the extension 2DPCA transform, we convert y to one-dimensional vector. The S_W and S_B are defined as

$$S_W = \frac{1}{L} \sum_{i=1}^c \sum_{j=1}^{L_{c_i}} (y_{i,j} - \mu_i)(y_{i,j} - \mu_i)^T \quad (15)$$

$$S_B = \frac{1}{L} \sum_{i=1}^c L_{c_i} (\mu_i - \mu)(\mu_i - \mu)^T \quad (16)$$

Where L_{c_i} , $y_{i,j}$ and μ_i are the number of feature vectors, the j th feature vector, and the mean vector in $c_i \in C$, respectively.

L denotes the total number of feature vectors in all classes,

and μ denotes the mean vector of all the feature vectors. The P_i in matrix $P = [P_1, P_2, \dots, P_m]^T$ is the generalized eigenvector of S_W and S_B corresponding to the i th largest eigenvalue λ_i .

$$S_B P_i = \lambda_i S_W P_i \quad (17)$$

Hence, after the extension 2DPCA and visemic LDA transforms, the final visual feature vector z can be obtained by

$$z = \tilde{P}y \quad (18)$$

4. Experiments

We now investigate the performance of the extension 2DPCA based feature extraction method. Before reporting the experiment results, we briefly describe the audio-visual database and the structure of our AVSR system.

4.1. The Audio-Visual Database

The CUAVE database [14] was considered in our experiments. This database consists of 36 speakers (19 male and 17 female) pronouncing the sequence of digits from "0" to "9", and the individual speakers are presented in the scene frontally facing the camera.

In our experiments, we first divide the database into a testing set and a training set. The training set consisted of 24 speakers (1200 utterances) and the testing set consisted of 12 speakers (600 utterances), the speakers in each of the sets were chosen randomly. The audio features are the 39 dimensional Mel Frequency Cepstral Coefficients (MFCCs) (12MFCCs, energy, first and second derivatives), which are extracted from a window of 25ms with an overlap of 15ms.

4.2. The AVSR System

Our audio-visual speech recognition system is built on the popular multi-stream hidden Markov model [3]. The combined multimodal observation stream, an audio stream O_t^A and a visual stream O_t^V ($0 < t < T$), can be represented as

$$O_t = [O_t^A, O_t^V], 0 < t < T \quad (19)$$

The log likelihood $b(O_t)$ of a combined vector O_t is defined by

$$b(O_t) = \lambda_A b_A(O_t^A) + \lambda_V b_V(O_t^V), \lambda_A + \lambda_V = 1, \lambda_A, \lambda_V \geq 0 \quad (20)$$

Where λ_A and λ_V are audio and visual stream weights, respectively. The multi-stream HMM model conducted by HTK HMM toolkit [15] is context-independent with 8 states per digit and a single Gaussian per state.

4.3. Result and Discussion

To evaluate the efficiency of presented visual feature extraction method, several experiments was conducted.

We first compared the word recognition accuracy rate of visual-only speech recognition by PCA, 2DPCA and the extension 2DPCA followed by LDA, respectively (Table 1). The number in the bracket (32×1 , 13) of the 'Dimension' column means that the dimension become 32×1 after PCA transform, and reduced to 13 after the LDA transform. Compared to 2DPCA, the extension 2DPCA reduces the dimension from 32×10 to 10×10 , and both of them have a similar recognition accuracy rate about 58%. From the table, we can see that the extension 2DPCA+LDA performs better than the PCA+LDA with about 3% increase. This can be explained from Fig3: when the extension 2DPCA uses 10 components for reconstruction, the Euclidean distance error

between the original test image and the reconstructed image is still lower than the value when PCA uses 32 components. On the other hand, considering the computational complexity of PCA and 2DPCA, clearly, the extension 2DPCA is more preferable.

Table1 The word recognition accuracy rate of the visual-only speech recognition using PCA, 2DPCA and extension 2DPCA followed by LDA.

Features	Recognition accuracy	Dimension
PCA+LDA	54.32%	(32×1,13)
2DPCA+LDA	58.16%	(32×10,13)
Extension 2DPCA+LDA	58.21%	(10×10,13)

In addition, in Table 2, we further compare the word accuracy rate of the audio-only and the extension 2DPCA based AVSR at different levels of SNR. Both audio-only speech recognition and AVSR achieve a high word recognition accuracy rate in a clean environment. When the SNR (signal-to-noise-ratio) is as low as 5dB, the AVSR system still has an accuracy rate about 67.57%, and the AVSR accuracy rate increases by about 43% over the audio-only speech recognition. Moreover, as we expected, the AVSR system significantly outperform the audio-only speech recognition when the SNR is reducing. It proves that in noisy environment AVSR system is more robust than audio-only speech recognition, which is also verified by many others' experiments in the literature too.

Table2 The word recognition accuracy rate (%) of the audio-only and audio-visual speech recognition at different SNR levels.

SNR(dB)	0	5	10	15	20	25	30
audio-only	9.26	24.32	54.34	76.72	89.91	90.64	92.36
AVSR	40.85	67.57	74.31	88.22	92.52	93.15	94.58

5. Summary

This paper presents an audio-visual speech recognition system by using the extension 2DPCA transform for visual feature extraction. In the post extraction process, the Visemic linear discriminant analysis is applied to form the final visual features. Experiments on a connected digits task demonstrated an obvious advantage of the proposed visual features over the traditional PCA transformed features. The extension 2DPCA can further reduce the dimension of 2DPCA. On the other hand, it can represent the testing image better than PCA does, and the extraction of visual features becomes more quickly and less computational than PCA because it does not need to transform the image matrix to one dimensional vector in advance. Moreover, the extension 2DPCA+LDA performs better than PCA+LDA in the visual-only speech recognition. Finally, the efficiency of the extension 2DPCA is identified in our AVSR system compared to audio-only speech recognition. However, we should note that these experiments are conducted in a small vocabulary database. Therefore, a context-dependent large vocabulary database should be considered in the future work.

6. References

- [1] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. PAMI.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.
- [2] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, Springer, Berlin, pp. 331–349, 1996.
- [3] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," *The Johns Hopkins University, Final Workshop 2000 Report*.
- [4] C. Bregler, Y. Konig, "Eigenlips" for robust speech recognition," in *Proc. ICASSP*, 1994.
- [5] G. Potamianos, A. Verma, C. Neti, G. Iyengar and S. Basu, "A cascade image transform for speaker independent automatic speechreading," in *proc. ICME*, 2000.
- [6] Lu Hong Liang, Xiao Xing Liu, Yibao Zhao, Xiaobo Pi and Ara V Nefian, "Speaker independent audio-visual continuous speech recognition", in *proc. ICME*, vol.2, p. 25-28, August 2002.
- [7] P. Duchnowski, M. Hunke, D. Biisching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," in *Proc. ICASSP*, 1995.
- [8] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. ICIP*, 1998.
- [9] J. Yang and J. Y. Yang, "From image vector to matrix: A straightforward image projection technique—IMPCA vs. PCA," *Pattern Recognition*, vol. 35, no. 9, pp. 1997–1999, Sep. 2002.
- [10] M. Visani, C. Garcia and C. Laurent, "Comparing Robustness of Two-Dimensional PCA and Eigenfaces for Face Recognition," *International Conference on Image Analysis and Recognition*, proto, 2004.
- [11] D. Xu, S. Yan, L. Zhang, M. Li, W. Ma, Z. Liu, H. Zhang, "Parallel image matrix compression for face recognition," *11th International Multimedia Modelling Conference*, pp. 232-238, 2005.
- [12] Wangmeng. Zuo, D. Zhang, J. Yang, K.Q. Wang, "BDPCA plus LDA: a novel fast feature extraction technique for face recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 36(4): 946-953, 2006.
- [13] C. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Reseach*, vol. 11, pp. 796-804, 1968.
- [14] E.K. Patterson, S. Gurbuz, Z.Tufekci, and J.N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. of ICASSP2002*.
- [15] S. Young, G. Evermann, et. al. "The HTK Book," <http://htk.eng.cam.ac.uk/>, Version 3.3, 2005.