A PCA based Visual DCT Feature Extraction Method for Lip-Reading

Xiaopeng Hong, Hongxun Yao, Yuqi Wan and Rong Chen
School of Computer Science and Technology, Harbin Institute of Technology, 150001, China
{xphong, yhx}@vilab.hit.edu.cn

Abstract

This paper proposes a PCA based method to reduce the dimensionality of DCT coefficients for visual only lip-reading systems. A three-stage pixel based visual front end is adopted. First, DCT or block-based DCT features are extracted. Second, Principal Component Analysis is applied for dimension reduction. Finally, all the feature vectors are normalized into a uniform scale. This work investigates this three-stage method, comparing with PCA and two DCT based approaches whose features are selected manually. In the latter manner, PCA coefficients are selected according to energy while the reduction of DCT coefficients leans to the left components in the left-top corner. Experiments prove that the dimension reduction task based on PCA does improve the recognition accuracy when the final dimension is below a certain value. They also show that DCT and block-based DCT work similarly for lip reading task, outperforming PCA slightly.

1. Introduction

Lip-reading has become a hot topic for human-computer interaction (HCI) and audio-visual speech recognition (AVSR). The speech recognition system which combines both auditory and visual information has been demonstrated to outperform the audio-only system [1, 3-4, 6, 8-9, 12, 14]. This paper concentrates on the visual-only lip-reading system which has also attracted significant interest [5, 11, 13, 15].

Feature extraction is a crucial part for a lip-reading system. Various visual features have been proposed in the literature. In general, feature extraction methods can be categorized into three kinds: 1. “pixel based” where features are employed directly from the image, 2. “lip contour based”, in which a prior template or model is used to describe the mouth area and 3. the combination of 1 and 2. Among these approaches, the one based on low level pixels is assumed to be the most efficient on [2, 4, 7, 14]. As a typical method to extract pixel based features, image transforms such as Discrete Cosine Transform (DCT) [7, 8, 10-12], Principal Component Analysis (PCA) [5-9,13], Discrete Wavelet Transform (DWT) [7] and Linear Discriminant Analysis (LDA) [10] have been employed for lip-reading and has achieved high accuracy for visual-only recognition task. Among these, DCT has been shown to perform equally well or better than others [14].

Working at this pixel-based field faces a problem: How to reduce the high dimensional raw image data to low dimensional feature vectors without losing important information? Potamianos [7] retained the coefficients according to several sub lattices. Heckmann [11] compared 3 strategies to select the coefficients based on energy, variance and relative variance respectively and stated that the one based on energy performed best. Nefian [12] divided a 64×64 Region of Interest (ROI) into 8 blocks of size 32×16, and extracted the first 2×2 low frequency coefficients from each block. Projection using LDA to seek optimal classification performance in [1, 15] can also be used for data dimensionality reduction, although this ability of LDA is limited to the number of classes [15].

Motivated by the above studies, this paper focuses on the dimensionality reduction strategies for DCT based features for visual-only lip-reading task. A new method is proposed in the current work. In view of the excellent ability for information compression, PCA is applied to extract DCT coefficients. This combination is assumed to utilize the advantages of these two transforms. DCT is preferable to differentiate frequencies while PCA is beneficial to select the most ‘important’ components. Experimental results demonstrate that this new method does improve the speech reading performance when the final dimension is below a certain point, compared to the methods selecting the coefficients according to specific criterion, such as ‘low frequency’.

This work also compares the effect of DCT with the block-based DCT.
This paper is structured as follows: A three-stage lip feature extraction method is discussed in Section 2. Three image transforms are described in Section 3. Section 4 focuses on the dimensionality reduction works and all the experiments are shown in Section 5.

2. Visual Feature Extraction Method

Before the visual feature extraction task, a region of interest (ROI) is centered at the speaker’s mouth area. Then they are down sampled to a 32 × 16 rectangle.

2.1. Overview

The visual feature extraction task concentrated in the current work is shown in Fig. 1. Inspired by the cascade strategy by [1], a three-step strategy, however different, is adopted here. The first stage is the image transform done in 3 styles respectively: DCT, block-based DCT and PCA. This step forms a 512 dimension vector (or a 32 × 16 matrix) T. The second step is the dimensionality reduction procedure, using PCA presented in this work or based on a specific criterion. The output of this step, a vector R, is normalized to form a final vector V in the final stage, similar to [10]. Then V is used as a feature vector for classification based on Semi-Continuous Hidden Markov Model (SCHMM) which is set with 6 states and 8 modes per state.

2.2. Normalization

The use of normalization in the last step mentioned above is due to the fact that coefficients produced by image transform always have different scales. Hence it is very difficult to simulate the distributions of them for GMM or SCHMM.

After the second stage, the dimension of R has been reduced to K. The following step calculates the mean μ and the standard deviation σ across all the vectors R_i, here
\[ R_i = \left[ r_{i,1}, r_{i,2}, \ldots, r_{i,K} \right]^T, \quad i = 1, 2, \ldots, S \]
(S is the number of training samples). R_i is then normalized to V_i using Eq. (1). After this process, the distribution of all V_i has a mean of 0 and a standard deviation of D/3, and what is more, 99.73% of the features have a uniform scale [–D, D].

\[ v_{i,j} = D \frac{r_{i,j} - \mu_j}{3\sigma_j}, \quad j = 1, 2, \ldots, K, \quad i = 1, 2, \ldots, S \]  
(1)

D is a free parameter chosen to adjust the scale of a feature vector. Note that D should be chosen according to the dimension of features. Training of HMM may fail with an improper D. D is set 2 in this work.

The same normalization is applied to the test data.

3. Image Transform

The intensity components from the down sampled 32 × 16 matrix in the ROI are transformed to form a vector T. Three pixel based features are investigated here: 1) PCA, 2) DCT and 3) 8 × 8 Block-based DCT.

3.1. Principal Component Analysis

According to the principle of minimizing mean square error, PCA is the optimal transform. It has been widely used for lip-reading task [5-9, 13]. In this manner, the 32 × 16 matrix is considered as a 1D vector X_i, i = 1, 2, ..., S (S has the same meaning as in Section 2.2). The mean vector m, the covariance matrix C for all X_i, the eigenvectors and eigenvalues of C are all calculated. For the sake of reducing the dimension, K eigenvectors with the largest eigenvalues are selected to generate the transform matrix P_{PCA}.

3.2. Two Kinds of DCT Features

In order to gain the DCT coefficients, there are two approaches: entire manner and block manner (Fig. 2). The front is more general, especially for a square ROI [7, 8, 10, 11] while Nefian [12] used latter. Both of these manners are investigated here to see whether the local characteristics are important for lip-reading task. Block-based DCT (labeled ‘B’ in Fig. 2) divides a 32 × 16 ROI image into 8 non-overlapping blocks of size 8 × 8 and applies DCT to each of the blocks to obtain the transform coefficients T_i^l, l = 1, 2, ..., 8.

All T_i^l are rearranged according to the specific order shown in (Fig. 2) to form T_i. For comparison, in the entire manner (labeled ‘E’ in Fig. 2), the whole 32 × 16 matrix is transformed. The experimental results
Section 5 show that these two approaches work similarly for visual-only lip-reading task.

![Fig. 2 Two strategies to gain the DCT coefficients. “B”: “Block-based DCT”. “E”: “Entire DCT”. “T”: “The output matrix”.

4. Feature Dimensionality Reduction

The objective of this stage is to reduce the dimension of $T$ without losing relevant information. As mentioned above, for PCA, the features corresponding to the $K$ largest eigenvalues are chosen manually.

Considering the strong ability for information compression, this work applies PCA to reduce dimension of the DCT based features. The matrix $T$ composing of DCT coefficients achieved by either entire manner or block manner is considered as a 512 dimension vector. This stage then outputs a $K$ dimension vector $R$: $R = P_{PCA} \times (T - m)$. Here $m$ is the mean vector for the training data which have been transformed by the first stage and $P_{PCA}$ is corresponding to the $K$ largest eigenvalues of the covariance matrix $C_t$ respect to the transformed training space.

![Fig. 3 The order of DCT coefficients in blocks]

In previous work [7, 11-12], the selection of the DCT coefficients was always based on some specific criterions. One criterion of this kind (Fig. 3) is investigated comparing with the method mentioned above. As a matter of fact, because the left components in the left-top corner are given a prior consideration, this order is strong to represent the characteristics of near-horizontal edges. Considering the shape of a lip, it is assumed to be beneficial. When in implementation, for entire DCT, the selection of coefficients obeys this order in a whole $32 \times 16$ matrix, while it is embodied in each block for the block-based DCT.

5. Result and Discussion

A database, Harbin Institute of Technology Bimodal Chinese Audio-Video Database (HIT Bi-CAVDB), was constructed, consisting of face images under natural luminance. It covers 96 different Chinese syllables (classes). All the images were at 25Hz and stored in 24bits, $256 \times 256$ image sequences. Some other configurations are depicted in Table. 1.

<table>
<thead>
<tr>
<th>Bi-CAVDB</th>
<th>Speakers</th>
<th>Subjects</th>
<th>Shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>200</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

All the experiments were carried out in Speaker Dependent (SD) manner. For each speaker, the first two shots of each subject were for training while the remaining was used for test.

First, Fig. 4 reported improved accuracy using PCA to reduce the dimension of DCT features. None of the manual selected features reaches accuracy over 70%. However, regardless the implementation manner of DCT, the PCA based reduction strategy gains a near 77% accuracy with a 10% increase over the manual selection method. The improvement is mainly because of the combination of two transforms. DCT coefficients are preferable to describe the appearance of the lip area, and then PCA gets rid of the redundancy of them. However, the improvement slow down when the dimension became higher, even fall at the point close to 80. This situation needs further investigation.

This work compared the effects of PCA and two DCT based methods with a manual selection. As shown in Table. 2, these three methods work similarly. All of them gained the highest accuracy up to 67% and below 70%. The dimension of PCA features is smaller due to its excellent ability for data representation. Two DCT based methods outperform PCA slightly. Considering the computational complexity of PCA, clearly, DCT based features are preferable. This result is in accordance with those obtained in [1, 2, 4].

Overall, the performance of DCT did not vary significantly in manual selected manner. The ituation is similar in PCA based selection (Fig. 4). When using manual selection, both of them reached over 68% accuracy (68.4% for Entire manner, 68.7% for Block manner). While using PCA reduction method, entire manner gained 77.1% and block manner 76.9%.
However, the block strategy is preferable for implementation in the non square 32x16 rectangle.

**Table. 2 Best results of pixel based features.**

<table>
<thead>
<tr>
<th>Feature</th>
<th>PCA</th>
<th>DCT</th>
<th>Block DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>45</td>
<td>64</td>
<td>80</td>
</tr>
<tr>
<td>Accuracy %</td>
<td>67.1</td>
<td>68.4</td>
<td>68.8</td>
</tr>
</tbody>
</table>

![Fig. 4 Result of Dimension reduction](image)

**6. Conclusions**

Combining advantages of the two transform, this paper proposes a PCA based method to reduce the dimension of DCT based features. This new strategy gains greatly improvements than the method based on specific criterion when the final dimension is below 80. Three image based features with manual selection are investigated here too. The DCT based features selected leaning to the left components in the left-top corner outperforms PCA slightly although all these strategies perform similarly. Experiments also prove that the block-based DCT works as efficient as the DCT for visual-only lip-reading system. In future work, the reason why the improvement of this PCA based dimensionality reduction approach slows down when the dimension closes to a certain point needs further researches.

**7. Acknowledgements**

This work is supported by: 1). Program for New Century Excellent Talents in University under the grant number NCET-05-03 34 and 2). Natural Science Foundation of Heilongjiang Province of China under the grant number E2005-29.

**8 References**