ABSTRACT

Talking face detection is important for videoconferencing. However, the detection of the talking face is difficult because of the low resolution of the capturing devices, the informal style of communication and the background sounds. In this paper, we present a novel method for finding the talking face using latent semantic indexing approach. We tested our method on a comprehensive set of home video conferencing sessions with a very high detection rate. Our experiments show that the LSI method accuracy degrades gracefully in a noisy environment as opposed to the correlation method which simply fails in presence of noise.

1. INTRODUCTION

Video communication has started to play an increasingly important role in our daily life. The quality of video conferencing experience can be improved by focusing on person talking at a given moment. This calls for automatic talking person detection and identification in video communication context. In the archived video communication sessions also, the availability of such information can be useful. The earlier related work on person identification has been concentrated on single mode processing, for example, face detection and recognition, speaker identification, and name spotting [1, 2]. A standard approach is to use fusion strategy or Bayesian network to integrate the results from speaker identification and face recognition. Also, speech recognition research has been recently focusing on audio-visual speech recognition where the visual analysis of lip movement is used to aid in recognizing words and raising overall accuracy. In both cases, it is assumed that the person appearing on the video is the person talking, which is not always true.

The work presented in this paper is aimed at finding the talking face in live or stored video and detecting the relationship between the speech and faces. We propose a mathematical framework that incorporates correlation and latent semantic indexing methods. One system was built to implement and test this framework. Our initial results show that the methods are promising.

This paper is organized as follows. The idea behind face-speech matching and representation of face and speech features are presented in Section 2. In Section 3, we give the techniques related to face-speech matching. The initial experimental results are presented in Section 4. The summary and future directions are discussed in Section 5.

2. TALKING FACE DETECTION

When a person is talking, he/she is always making certain facial and head movements. In the normal manner of speech, the head is turning right, left, back, and forth, speaker’s mouth is opening and closing. In most cases, the person is making different facial expressions and gestures. In our work, we focus on head movement and face changes and how these movements affect the face subimage.

We have two means to capture mouth movement. First, we can track the movement of mouth. A lot of work in speech recognition is done on lip reading, where tracking of lip movement helps to determine which word is pronounced. However, in lower resolution video, it is not easy to track the lips’ movement. Instead, we can track face change resulting from lip movement. As we know, with the lip movement, the color intensity of lower face image will change, face image size will change slightly too. Through tracking change in the lower part of face image, we can coarsely track the lip movement. Change in facial expression while talking will also have an effect on changing face image. For this reason, our system tracks the position of head and lower image of face, which reflects the movement of both the head and the lips.

2.1 Feature Representation
For the face image, we use the principal component analysis (PCA) representation [9], which reduces the number of features dramatically. The sensitivity of PCA to face direction is useful for us. In PCA, the eigenvectors represent global feature of face.

For the audio part, we use 20 audio features consisting of average energy, pitch, zero crossing, bandwidth, bandcentral, roll off, low ratio, spectral flux and 12 MFCC components [6]. We use K audio features to represent the speech signal. Thus for each frame, we have a K dimensional vector to represent speech in this frame as

\[ A = (a_1, a_2, \ldots, a_K) \]

where ' represents matrix transposition.

For each face, we use I features to represent it. So for each frame, we have I dimension vector for each face. Assume there are M faces in the video the feature vectors for faces for each frame will be as follows.

\[ F = (f_1^1, f_2^1, \ldots, f_I^1, f_1^2, f_2^2, \ldots, f_I^2, \ldots, f_1^K, f_2^K) \]

We first list all the components of visual face features, then audio features. The resulting vector is

\[ V = (f_1^1, f_2^1, \ldots, f_I^1, f_1^2, f_2^2, \ldots, f_I^2, \ldots, f_1^K, f_2^K, \ldots, f_K^K) \]

which represents all the information about the face and speech in one frame. Assume there are N frames in one trajectory, an (IM+K)xN matrix represents this trajectory where the i-th column represents i-th frame V vector.

3. TECHNIQUES FOR FACE SPEECH MATCHING

We use two methods to compute the face-speech matching. The first one is correlation method. The second method is LSI (Latent Semantic Indexing). LSI is a powerful method in text information retrieval that uncovers the inherent and semantic relationship between different objects. We use it here to find the inherent relationship between spoken audio and corresponding faces.

3.1 Correlation method

For the correlation method, we calculate the correlation between audio vector and face vectors and select the face having maximum correlation with audio. The rationale is that there is an inherent relation between the speech and speaking person. We compute the correlation between the audio and speech vectors according to the following well-known equations. First, the mean vector of the video is given by:

\[ V_m = \sum_{i=1}^{N} V_i \]

The covariance matrix of V is given by:

\[ C(i,j) = \frac{\hat{C}(i,j)}{\sqrt{\hat{C}(i,i)\hat{C}(j,j)}} \]

The normalized covariance is given by

\[ \tilde{C} = \sum_{i=1}^{N} (V_i - V_m)(V_i - V_m)^\prime \]

The correlation matrix between the speech vector A and the m-th face vector is the submatrix C(IM+1:IM+K, (m-1):m). We can compute the sum of all the elements of this submatrix, denoted as c(m), which is the correlation between m-th face vector and the audio vector. Then we choose the face having the maximum correlation.

3.2 Latent Semantic Indexing Method

The LSI uses singular value decomposition (SVD) in matrix computations to obtain a new representation for two sets of entities, for example keywords and documents in information retrieval applications. The new representation uses uncorrelated basis vectors and allows for approximation or dimension reduction. In this way, LSI achieves three goals: dimension reduction, noise removal, and uncovering of the semantic and hidden relation between different objects, like keywords and documents.

In our current context, we can use LSI to relate the low level representation of image and audio features in a video sequence. First we need to build a matrix for the video sequence using the frame vectors introduced in the previous section as:

\[ \hat{X} = (V_1, V_2, \ldots, V_N) \]

Since each component of V is heterogeneous consisting of the visual and audio features:

\[ V = (f_1^1, f_2^1, \ldots, f_I^1, f_1^2, f_2^2, \ldots, f_I^2, \ldots, f_1^K, f_2^K, \ldots, f_K^K) \]

we first normalize each component by their maximum elements as:

\[ X(i,:) = \frac{\hat{X}(i,:)}{\max(\text{abs}(\hat{X}(i,:)))} \]

where X(i,:) denotes the i-th row of matrix X, and the denominator is the maximum absolute element of i-th row. The resulting matrix X has elements between -1 and 1. Assume the dimension of V is H, then X is a HxN matrix.

We perform singular value decomposition on X as follows:

\[ \hat{X} = SVD' \]

Where S is composed of the eigenvectors of XX' column by column, D consists of the eigenvectors of X'X, \( V^2 \) is a diagonal matrix where diagonal elements are eigenvalues.
Normally, the $S$, $V$, and $D$ matrices must all be of full rank. The strength of SVD, however, lies in that it allows a simple strategy for optimal approximate fit using smaller matrices. Let us order the eigenvalues in $V$ in descending order and keep the first $k$ elements, then we can represent $X$ by

$$X \approx \hat{X} = \hat{S}\hat{V}\hat{D}'$$

Where $\hat{V}$ consists the first $k$ elements of $V$, $\hat{S}$ consists the first $k$ columns of $S$ and $\hat{D}$ consists the first $k$ columns of $D$. It can be shown that $\hat{X}$ is the optimal representation of $X$ in least square sense.

After having the new representation of $X$, we can perform various operations in the new space. For example, we can compute the correlation of the face vector and the audio vector, compute the distance between face vector and the audio vector, and compute the difference between frames to perform frame clustering. For our face-speech matching, we compute the correlation between face features and audio features in $\hat{X}$ as explained in section 3.1. How to choose $k$ is an open issue. Here we choose $k$ so that it is large enough to keep the main information of the underlying data and at the same time small enough to remove noise and unrelated information.

4. EXPERIMENTAL RESULTS

We performed a series of experiments to test our ideas and methods. We used different types of video material to test our data: conferencing session, three discussion group sessions, and instructional video (Video Scout). The total duration of video in the experimental data set is 66 minutes. In all, we can classify the video into three categories:

- Good quality: low noise level, no large, frequent head movements
- Low quality: high noise level, two speakers’ voice almost can not be heard,
- Normal quality: low noise level, small movements.

Our experiments consist of two parts. The first one is used to illustrate the relationship between audio and video. The second part illustrates the test face-speech matching.

4.1 Illustration of audio and video relation

The general relationship between audio and video was tested by Wang et al. [7]. They computed the relationship between audio features with global video frame features, which showed weak or no relationship between audio and video frames. However, the global features used in their experiment carries too much noise hiding the relationship between audio and video. Thus, we just use the facial features to calculate the relationship. We present two methods to compute the correlation between audio and video:

- Straightforward correlation between the audio and video as explained in section 3.1.
- Correlation between the audio and video in the new space transformed by LSI as explained in section 3.2.

Figure 1 Correlation between audio and corresponding speaking face (left) and another non-speaking face (right).

First, we calculate the normalized correlation between faces and speech directly. The correlation matrix is shown in Fig. 1 (left). Each cell represents the corresponding element of matrix: the bigger the element, the whiter the cell. The figure 1 (left) represents the correlation matrix for a talking face, which reflects the relationship between speaker’s face with his voice. Fig. 1 (right) shows the correlation matrix between a silent listener with the same speech. The first four elements are eigenvalues of the talking face, the remaining elements are audio features: average energy, pitch, zero crossing, bandwidth, bandcentral, roll off, low ratio, spectral flux and 12 MFCC components, respectively. The elements in the four columns under 4th row in Fig. 1 (left) are much brighter: the speakers’ face has relation with his voice: sum of these elements in Fig. 1 (left) is 15.6; the sum of these elements in Fig. 1 (right) is 9.9.

Figure 2. The relationship between the average energy (dash-dot) with the speaker’s eigenface (solid) and listener’s eigenface (dotted).
The first four columns of the 5th row and 6th row in Fig. 1 (left) are much brighter than in Fig. 1 (right). The sum of these eight elements is 3.5 in Fig. 1 (left), and 0.7 in Fig. 1 (right). The 5th row represents the correlation between face and the average energy. The 6th row represents the correlation between face and pitch. When a person is talking, his face is changing too. Voice’s energy is corresponding to the opening and closing of the mouth. To show their relationship, we show the speaker’s and listener’s eigenface and average energy with time in Fig. 2. The dash-dotted line represents the average energy; the solid line represents speaker’s first eigenface; the dotted line depicts the listener’s first eigenface. We see that the eigenfaces are changing with average energy, since they have the same trend, while the listener’s face does not change.

We also compute the correlation of audio and video features in the new space transformed using LSI. A sample result is shown in Fig. 3. The first two components are the speaker’s eigenfaces, and the next two components are the listener’s eigenfaces; the next components are audio features. From this, we see that the first two columns are brighter than the next two columns, which means that speaker’s face is correlated with his voice.

![Eigen features and Audio features](image)

**Figure 3. An example correlation matrix after LSI.**

### 4.2 Face-speech matching

Table 1 presents the whole set with different types of data. The duration is shown in number of minutes.

<table>
<thead>
<tr>
<th>Type</th>
<th>Normal</th>
<th>Good</th>
<th>Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>35</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>Persons</td>
<td>9</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td># of voices</td>
<td>9</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table I. Description of the data set of video sessions.

First, we compute the correlation between audio features and eigenfaces for each face-speech pair according to Section 3.1. We choose the face as speaker, which has maximum correlation with audio. Then we perform LSI on each pair first, and compute the correlation between audio and face features. The results are given in Table II. What is immediately obvious from the experimental results is that LSI proved its power in noise removal particularly for noisy data. On the other hand, the results show that the raw correlation shows reasonable performance for good data, however is not usable in the presence of noise.

<table>
<thead>
<tr>
<th>Type</th>
<th>LSI</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Data</td>
<td>75%</td>
<td>57%</td>
</tr>
<tr>
<td>Good Data</td>
<td>80%</td>
<td>68%</td>
</tr>
<tr>
<td>Noise Data</td>
<td>66%</td>
<td>34%</td>
</tr>
</tbody>
</table>

Table II. Accuracy of speaking face detection.

### 5. SUMMARY DISCUSSION

In this paper, we suggested a framework to find the talking face in live or stored video by face-speech matching. Our experiments show that using the LSI approach is more promising even in presence of noise. The LSI method on good data can have 80% accuracy as opposed to 68% for the correlation method. In presence of noise LSI accuracy gracefully degrades to 66% while the correlation method slumps down to 34%.

For the future work, there are three directions. The first one is to continue research on face-speech matching and concentrate on feature selection and experimenting with a wider range of test video clips. The second direction is to include more information sources, such as transcript. The third direction is to extend face-speech matching to video understanding, build association between sound and objects that exhibit some kind of intrinsic motion while making that sound.

### 6. References