AUTOMATIC EXTRACTION OF GEOMETRIC LIP FEATURES WITH APPLICATION TO MULTI-MODAL SPEAKER IDENTIFICATION

Ivana Arsic, Roger Vilagut and Jean-Philippe Thiran

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email: {Ivana.Arsic, Roger.Vilagut, JP.Thiran}@epfl.ch
http://itswww.ep.ch

ABSTRACT
In this paper we consider the problem of automatic extraction of the geometric lip features for the purposes of multi-modal speaker identification. The use of visual information from the mouth region can be of great importance for improving the speaker identification system performance in noisy conditions. We propose a novel method for automated lip features extraction that utilizes color space transformation and a fuzzy-based c-means clustering technique. Using the obtained visual cues closed-set audio-visual speaker identification experiments are performed on the CUAVE database, [1] showing promising results.

1. INTRODUCTION
An exhaustive research done in the area of joint audio-visual speech processing over the last decade shows that the usage of visual modality can be beneficial for improving the overall system performance, especially in the cases when the environmental conditions are changed. The inclusion of visual modality together with audio have been already used in various applications such as: audio-visual speech recognition [2], bimodal speaker recognition [3], speaker localization, etc.

The main problem when trying to benefit from speech and speaker dependent information conveyed in the lip movement, is the extraction of the suitable lip features. Moreover, the applied method should be robust and accurate, and if possible accomplished in an automatic manner. Various techniques were proposed by different researchers and an excellent overview of different methods can be found in [2].

In this paper we address the problem of automatic extraction of geometric lip features, namely four outer lip contour points from the observed lip region images. Further on, the obtained geometric lip descriptors are applied in the framework of audio-visual speaker identification on the CUAVE database [1], previously used mainly for audio-visual speech recognition. The recent work by Dean et al. [4], is the first showing that this audio-visual corpora can be equally used to test speaker identification performance using eigenlips as visual cues. This motivated us to test the possibility of using geometric based features for the same tasks.

Section 2 describes an algorithm for extraction of geometric lip features. In Section 3 we recall the mathematical background regarding the speaker modeling using Gaussian mixture models. The experimental framework and obtained results are presented in Section 4, followed by the conclusions in Section 5.

2. VISUAL FEATURE EXTRACTION
The extraction of visual features is performed in two stages. The first considers the lip Region-of-Interest (ROI) localization in each frame of the audio-visual sequence. In the second stage a lip segmentation technique based on fuzzy c-means clustering is applied on the lip ROI and features are extracted.

2.1. Lip ROI extraction
Prior to the lip Region-of-Interest (ROI) extraction, a face detection is performed using the algorithm described in [5]. As a result, the square bounding box and a center of the subject’s face is found. Then we use the template matching of the nose region between consecutive frames in order to stabilize the detection process. Knowing the nose (face) center position, we use anthropometrical properties of the human face to define a region around the center of the mouth from the first frame. Subsequently, from each frame of size 360 × 240 of the observed video sequence the 60 × 50 pixels RGB color lip ROI is localized and extracted, as shown in Figure 1.

2.2. Color space transformation
As a first step a color transformation from the RGB color space to CIELAB and CIELUV is performed on the extracted lip ROI. The main idea is to obtain perceptually uniform color spaces, as well as to increase robustness to the lip/skin color...
For each algorithm based on fuzzy c-means clustering, similar to the one presented in [7]. The lip features of interest are four outer lip contour points, namely lip corners and upper and lower lip middle points. The extraction procedure is performed through an iterative process in the following three steps, repeated until convergence: masks and centroids calculation, membership function calculation, and feature extraction and error correction.

2.4. Mask and centroids calculation

For each component image an initial set of masks \( M_i \) is built using the color distribution information from the image histogram and morphological filtering. Color centroids are calculated by overlaying each component mask to the corresponding component image. Hence, we can calculate color centroids for lip region \( v_{\text{lip}} \) and skin region, \( v_{\text{skin}} \), as following:

\[
v_{\text{lip},i} = \frac{1}{l_i} \sum_{r,c} x^*(r,c) \cdot M_i(r,c)
\]

\[
v_{\text{skin},i} = \frac{1}{m \cdot n - l_i} \sum_{r,c} x^*(r,c) \cdot (1 - M_i(r,c)) \tag{1}
\]

where \( x^*(r,c) \in \{a^*, b^*, u^*, v^*\} \) represents a pixel in the row \( r \) and column \( c \) in the image of dimension \( m \times n \). Parameter \( l_i \) denotes the size of the mask, i.e. the number of pixels having the intensity equal to 1 for the lip region, and 0 for the skin region.

2.5. Membership function calculation

The next step is a membership function calculation and the objective is to maximize the dissimilarity between pixels from different clusters.

Let \( d_{i}(r,c) \in \{d_{i,a}(r,c), d_{i,b}(r,c), d_{i,u}(r,c), d_{i,v}(r,c)\} \) stand for the Euclidean distance between the feature vector of each pixel \( x(r,c) \) and the color centroids \( v_{\text{lip}} \) and \( v_{\text{skin}} \). We calculate the color dissimilarity \( D_i(r,c) \) for lip and skin as:

\[
D_{\text{lip}}(r,c) = d_{\text{lip},a}(r,c) + d_{\text{lip},u}(r,c) + \frac{1}{3}(d_{\text{lip},b}(r,c) + d_{\text{lip},v}(r,c)) \tag{2}
\]

\[
D_{\text{skin}}(r,c) = d_{\text{skin},a}(r,c) + d_{\text{skin},u}(r,c) + \frac{1}{3}(d_{\text{skin},b}(r,c) + d_{\text{skin},v}(r,c)) \tag{3}
\]

In equations 2 and 3 the experimentally obtained weighting parameter \( \frac{1}{4} \) is introduced in order to decrease the influence of masks \( M_b \) and \( M_v \) since they can be of a poor quality. Finally, the membership map for the lip cluster is calculated as:

\[
u_{\text{lip}}(r,c) = \frac{D_{\text{skin}}(r,c)}{D_{\text{lip}}(r,c) + D_{\text{skin}}(r,c)} \tag{4}
\]

The resulting membership map of an example lip ROI at each iteration is shown in Figure 3 (a). In order to segment the lip area, the probability map is thresholded at 0.5 and the biggest connected area represents the lips. The undesired background effects can be eliminated by utilizing the lip symmetry. Segmented lip images after thresholding for different iteration steps are shown in Figure 3 (b).

2.6. Lip features extraction

After locating the lip area in the lip ROI, we can estimate coordinates of the lip features of interest. The left and right lip corners are found from the \( x \) and \( y \) image intensity profiles. Then, the mouth center coordinates are calculated by finding the mean value of the distance between the lip corners. An orthogonal to the line connecting the lip corners is
As a final step, temporal smoothing using a median filter is applied to the extracted set of lip coordinates. Some example image frames of different speakers and located lip points after running the algorithm are shown in Figure 4.

3. SPEAKER MODELING

The baseline of our speaker identification system is built using a state-of-the-art Gaussian mixture model approach (GMM) as in [8]. One GMM model is built for each speaker from the database, representing all the speech utterances of that speaker. Thus, a speaker \( S \) is represented by a model \( \lambda_S \):

\[
\lambda_S = \{ p_i^S, \mu_i^S, \Sigma_i^S \} \quad \text{with} \quad i = 1, \ldots, M. \quad (6)
\]

where \( p_i^S \), \( \mu_i^S \) and \( \Sigma_i^S \) are the mixture weight, mean vector and covariance matrix for the mixture \( i \) from the set of \( M \) mixture components. The goal of the speaker identification tasks is to find the index \( S \) of the speaker from the set of speakers \( S \) whose model will have the maximum a posteriori probability given the observation sequence \( O \):

\[
\hat{S} = \arg \max_{1 \leq k \leq S} \{ p(O) | \lambda_k \} = \arg \max_{1 \leq k \leq S} \prod_{t=1}^{T} p(o(t) | \lambda_k) \quad (7)
\]

4. EXPERIMENTS AND RESULTS

The lip feature extraction algorithm is tested on audio-visual sequences form the Clemson University database CUAVE, [1]. This database consists of 36 speakers having different skin and lip tones, with various visual occlusions such as: glasses, beards and hats. We consider only part of the database where an individual speaker is present in the scene frontally facing the camera, pronouncing the sequence of digits from “zero” to “nine”, five times in repetition. For the majority of the video clips used (33 out of 36), the method performed well. In certain cases the problem appeared either due to the thin lips of the subject and/or significant head movement. For these sequences lip points of interest were marked manually every 10 frames.

After implementing the previously described algorithm, width and height of the lips are extracted from each frame of the video clip and used to form a visual feature vector.

4.1. Speaker identification results

Using the obtained visual features we perform visual-only speaker identification experiments. For each trial, the training set consists of the four repetitions of the sequence of digits from “zero” to “nine”, while the remaining fifth sequence...
is used for testing. The identification results for each test trial are represented in Table 1. The best overall results are obtained with number of Gaussian mixtures between three and five.

<table>
<thead>
<tr>
<th>test sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>scores [%]</td>
<td>77.78</td>
<td>88.89</td>
<td>94.44</td>
<td>94.44</td>
<td>83.33</td>
</tr>
</tbody>
</table>

**Table 1.** Visual-only speaker identification scores for five test sequences.

Further on, we perform audio-only and audio-visual identification experiments. Acoustic features are represented with 12 Mel Frequency Cepstral Coefficients (MFCCs) commonly used features in speech recognition. Tests were done for both clean and degraded audio conditions. Noisy environment is simulated by adding white Gaussian noise to the speech signal at different SNR levels ranging from 30 dB to -6 dB in steps of 6 dB. All models were trained using clean speech and tested on noisy data.

For audio-visual experiments we use the composite audio-visual feature vector i.e. feature fusion strategy. In order to have the same sampling rate for both audio and visual data streams, the audio is downsampled to the rate of video, or equivalently 30 Hz. Identification scores when using joint audio-visual feature vector are given in Table 2 for both clean and noisy conditions. It can be seen that the help of visual modality significantly improves the system performance when compared to audio-only identification results in noisy conditions. However, the overall audio-visual speaker identification system performance is far from satisfying. The lower scores than the video-only are due to the chosen feature fusion strategy and audio and visual streams misalignment.

<table>
<thead>
<tr>
<th></th>
<th>audio</th>
<th>visual</th>
<th>audio-visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>100</td>
<td>83.33</td>
<td>100</td>
</tr>
<tr>
<td>30 dB</td>
<td>91.67</td>
<td>83.33</td>
<td>97.22</td>
</tr>
<tr>
<td>24 dB</td>
<td>93.94</td>
<td>83.33</td>
<td>94.44</td>
</tr>
<tr>
<td>18 dB</td>
<td>50.00</td>
<td>83.33</td>
<td>77.78</td>
</tr>
<tr>
<td>12 dB</td>
<td>19.44</td>
<td>83.33</td>
<td>61.11</td>
</tr>
<tr>
<td>6 dB</td>
<td>8.33</td>
<td>83.33</td>
<td>41.67</td>
</tr>
<tr>
<td>0 dB</td>
<td>5.56</td>
<td>83.33</td>
<td>30.56</td>
</tr>
<tr>
<td>-6 dB</td>
<td>2.78</td>
<td>83.33</td>
<td>27.78</td>
</tr>
</tbody>
</table>

**Table 2.** Speaker identification results using feature fusion and test sequence #5.

### 5. CONCLUSION

We present a novel approach for locating the lip corners, as well as the upper and the lower outer lip contour middle points, using color transformation and a modified fuzzy c-means clustering algorithm. The proposed method is tested on 36 sequences from the CUAVE database working well for the majority of tested speakers. The coordinates of four key lip feature points are used to estimate the width and height of the speakers’ mouth. Then we perform various audio-visual speaker identification experiments. Our preliminary tests done on the CUAVE database show promising results in terms of utilizing the geometric lip features, for the identification tasks on the data corpus, previously used mainly for audio-visual speech recognition.

Future work would be to test the lip feature extraction method in the presence of video noise, as well as to perform audio-visual speaker identification experiments using other visual cues. Since the audio-visual identification scores when using early integration fusion method are still far from satisfying in a noisy environment, the other possible research direction would be towards utilizing adaptive decision fusion methods.

### 6. REFERENCES


