Combining Dynamic Texture and Structural Features for Speaker Identification

Guoying Zhao
Machine Vision Group
Infotech Oulu and Department of Electrical and Information Engineering
P. O. Box 4500 FI-90014
University of Oulu, Finland
gyzhao@ee.oulu.fi

Xiaohua Huang
Machine Vision Group
Infotech Oulu and Department of Electrical and Information Engineering
P. O. Box 4500 FI-90014
University of Oulu, Finland
huang.xiaohua@ee.oulu.fi

Yulia Gizatdinova
Department of Computer Science,
Tampere University, Finland
yulia.gizatdinova@cs.uta.fi

Matti Pietikäinen
Machine Vision Group
Infotech Oulu and Department of Electrical and Information Engineering
P. O. Box 4500 FI-90014
University of Oulu, Finland
mkp@ee.oulu.fi

ABSTRACT

Visual information from captured video is important for speaker identification under noisy conditions that have background noise or cross talk among speakers. In this paper, we propose local spatiotemporal descriptors to represent and recognize speakers based solely on visual features. Spatiotemporal dynamic texture features of local binary patterns extracted from localized mouth regions are used for describing motion information in utterances, which can capture the spatial and temporal transition characteristics. Structural edge map features are extracted from the image frames for representing appearance characteristics. Combination of dynamic texture and structural features takes both motion and appearance together into account, providing the description ability for spatiotemporal development in speech. In our experiments on BANCA and XM2VTS databases, the proposed method obtained promising recognition results comparing to the other features.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology, Theory and methods

General Terms
Algorithms

Keywords
Speaker identification, dynamic visual features, local spatiotemporal descriptors.

1. INTRODUCTION

Speaker identification is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals and lip movements. Comprehensive reviews of speech and speaker recognition can be found in [9]. A new review concerning audio-visual biometrics combining voice, visual speech and face is presented in [1]. Most of the early researches focused on using audio signal for this task [4, 7, 9, 12, 15, 24]. The dynamic visual features are suggested based on the shape and intensity of the lip region because changes in the mouth shape including the lips and tongue carry significant phoneme-discrimination information. An intuitive understanding of features related to lip movement is that different people have different articulatory styles when speaking the same utterance. For instance, when uttering the same phoneme or word, some speakers tend to widely open the mouth while others may only keep slightly opened; some speakers tend to move the lips in a specific direction, and for some speakers, the teeth are always visible when uttering [18].

In recent years, some techniques have been suggested that combine visual features with audio features to improve the recognition [5, 6, 8, 11, 21]. Audio features are still the main part and play more important role. However, in some cases, it is difficult to extract useful information from the audio. Detecting a person’s speech from a distance or through a glass window, or among a very noisy crowd of people, are some examples. In these applications, the performance of traditional speaker recognition is very limited. There are many applications in which it is necessary to recognize speaker using sole visual information in acoustically noisy environments that have background noise or cross talk.
among speakers. There are a few works focusing on the lip movement representations for speaker recognition solely with visual information [2, 3, 13, 18]. In [18], mouth corner is first detected, then static geometric features and dynamic lip movement features are investigated. Cetingul et al. [2] proposed to use 2D discrete cosine transform (DCT) coefficient to describe lip motion and further analyzed by Linear Discriminant Analysis (LDA) method to reduce the dimension. Based on this, they combined the 2D-DCT coefficients of the motion vectors within the mouth region and the 1D-DCT motion vectors around the shape of the lip contour to represent the lip motion. Luetten et al. [13] proposed lip boundary and intensity parameters which describe the grey-level distribution of the mouth area for speaker identification.

It can be seen that there are two kinds of commonly used features for describing the lip motion. One is geometric feature, including fiducial points like facial animation parameters, contours of lips and so on. These methods commonly require accurate and reliable facial and lip feature detection and tracking, which are very difficult to accommodate in practice and even impossible at low image resolution.

A desirable alternative is to extract features from the gray-level data directly. This type of features is based on observing the whole mouth Region-of-Interest (ROI) as visual informative about the spoken utterance. The feature vectors are computed using all the video pixels within the ROI. The proposed approaches include DCT or other transforms.

It appears that most of the research on visual speech recognition based on the ROI features, like DCT, has considered global features of lip or mouth images, but omitting the local features. Local features can describe the local changes of mouth in space and time. In this paper, we focus on the speaker recognition using only visual information. A new representation combining local dynamic texture and structural features is proposed for motion and appearance description, taking into account the motion of mouth region and the transition in appearance and utterance motions.

2. LOCAL SPATIOTEMPORAL DESCRIPTORS

2.1 Dynamic Texture Features for Motion Representation

The local binary pattern (LBP) operator is a gray-scale invariant texture primitive statistic, which has shown excellent performance in the classification of various kinds of textures [17]. For each pixel in an image, a binary code is produced by thresholding its neighborhood with the value of the center pixel (Fig. 1 and Eq. 1).

\[
LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}
\]  

where, \( g_c \) corresponds to the gray value of the center pixel \((x_c, y_c)\) of the local neighborhood and \( g_p \) to the gray values of \( P \) equally spaced pixels on a circle of radius \( R \). By considering simply the signs of the differences between the values of neighborhood and the center pixel instead of their exact values, LBP achieves invariance with respect to the scaling of the gray scale.

A histogram is created to collect up the occurrences of different binary patterns. The definition of neighbors can be extended to include circular neighborhoods with any number of pixels. In this way, one can collect larger-scale texture primitives or micro-patterns, like lines, spots and corners [17]. “Uniform patterns” [17] are usually used to shorten the length of the feature vector of LBP.

Recently, a method for temporal texture recognition using spatiotemporal local binary patterns extracted from three orthogonal planes (LBP-TOP) was proposed [23]. With this approach the ordinary LBP for static images was extended to spatiotemporal domain and region-concatenated descriptors using LBP-TOP features were developed for facial expression recognition [23] and visual speech recognition [22]. The obtained results outperformed the state-of-the-art. The LBP features extracted from XT and YT planes are utilized to represent the motion of utterance by catching the transition information of micro textures for speaker identification in this paper.

2.2 Structural Features for Appearance Representation

Recently, edge map (EdgeMap) features were proposed to describe the edge orientation of the pixel for detecting facial landmarks [10]. Since they can catch the structural information well, we use them to describe the appearance of the utterance.

Local oriented edges are extracted by convolving a smoothed image with a set of \( K \) kernels, as Fig. 2 shows. Each kernel is sensitive to one of the \( K \) orientations. The whole set of \( K \) kernels results from differences between two oriented Gaussians with shifted kernels (Eq. 2).

\[
G_{\phi_k} = \frac{1}{Z} (G_{\phi_k}^+ - G_{\phi_k}^-) \\
Z = \sum_{p,q} (G_{\phi_k}^- - G_{\phi_k}^+), G_{\phi_k}^- - G_{\phi_k}^+ > 0 \\
G_{\phi_k}^+ = \frac{1}{2\pi\sigma^2} \exp(- \frac{(p - \sigma \cos \phi_k)^2 + (q - \sigma \sin \phi_k)^2}{2\sigma^2}) \\
G_{\phi_k}^- = \frac{1}{2\pi\sigma^2} \exp(- \frac{(p + \sigma \cos \phi_k)^2 + (q + \sigma \sin \phi_k)^2}{2\sigma^2})
\]

where \( \sigma \) is a root mean square deviation of the Gaussian distribution, \( \phi_k \) is angle of the Gaussian rotation, \( \phi_k = 360/K \times k, k = 0, \ldots, K - 1 \); \( p,q = -P, -P + 1, \ldots, P - 1 \).

The maximum response of all \( K \) kernels (Eq. 6) defines the contrast magnitude of a local edge at its pixel location. The orientation of a local edge is estimated with the orientation of a kernel that gave the maximum response.

\[
g_{ij \phi_k} = \sum_{p,q} b_{i-p,j-q} G_{\phi_k}
\]

where \( b \) denotes the gray level of the image at pixel \((i,j)\); \( i = P \cdots W - 1 - P, j = P \cdots H - 1 - P; W, H \) are, respectively, the width and height of the image. 

\[
\text{Figure 1: LBP operator.}
\]
Figure 2: Orientation template with $\varphi_k = 360/K \times k, K = 16$.

Figure 3: Histogram of edge orientations (EdgeMap) with $K = 10$.

After getting the edge orientation for each pixel, a histogram is created to collect up the occurrences of different orientations, which is called EdgeMap, as Fig. 3 shows. The occurrence of the edge orientations of the pixels in the image is used to describe the structural feature.

2.3 Combined Dynamic Texture and Structural Features

We combine the structural EdgeMap features from the XY plane to describe the appearance and the local binary features from the temporal XT and YT planes to describe the motion.

The descriptors are obtained by concatenating the EdgeMap features from XY plane and local binary patterns from XT and YT planes. Fig. 4 (a) demonstrates the volume of utterance sequence. (b) shows image in the XY plane. (c) is image in XT plane providing visual impression of one row changing in time, while (d) describes the motion of one column in temporal space. A description computed over the whole utterance sequence encodes only the occurrences of the micro-patterns without any indication about their locations. To overcome this effect, a representation which consists of dividing the mouth image into several overlapping blocks is introduced. Fig. 4 also gives some examples of the EdgeMap, LBP image sequences, which are drawn using EdgeMap from XY (e), LBP code of every pixel from XT (f) and YT (g) planes, respectively, corresponding to transition information in appearance (b), horizontal motion (c) and vertical motion (d). From this figure, the changes in appearance and motion during utterance can be seen.

The EdgeMap histograms from XY plane and LBP histograms from XT and YT planes in each block volume (Fig. 4 (h)) are computed and concatenated into a single histogram, as Fig. 4 (i) and (j) shows. All features extracted from each block volume are connected to represent the appearance and motion of the mouth region sequence.

In this way, we effectively have a description of the utterance on three different levels of locality. The labels (bins) in the histogram contain structure information from XY plane, dynamic texture information from XT and YT planes, describing appearance and temporal information at the pixel level. The labels are summed over a small block to produce information on a regional level expressing the characteristics for the local appearance and motion in specific locations and time segment, and all information from the regional level is concatenated to build a global description of the mouth region motion.

Our method consists of three stages. The first stage is to detect the face, and then the eyes. The positions of the eyes are used to localize the mouth region. The second stage extracts the visual features from the mouth movement sequence. The role of the last stage is to recognize speakers from the input utterance using a classifier.

3. EXPERIMENTS

Experiments on two databases were carried out for evaluating the performance of the proposed method: the BANCA database and XM2VTS database.

Boosted Haar features [20] are used for automatic coarse face detection and 2D Cascaded AdaBoost [16] is applied for localizing eyes in the detected faces. Because the face images in the database are of good quality and almost all of them are frontal faces, detection of faces and eyes is quite easy. The positions of the two eyes in the first frame of each sequence were given by the eye detector automatically and then these positions were used to determine the fine facial area and localize the mouth region using some predefined ratio parameters [22] for the whole sequence. Fig. 5 demonstrates the input image sequences and the obtained mouth region images.

For feature extraction, each mouth image is divided into $1 \times 5$ blocks. From each block volume, the EdgeMap features with $K = 60$ were extracted from XY plane, and uniform LBP features with eight neighboring points and radius three were computed from XT and YT planes. The concatenated
spatiotemporal EdgeMap-LBP features are then utilized as inputs to the classifier.

3.1 BANCA Database

The BANCA database (http://www.ee.surrey.ac.uk/CVSSP/banca/) is a large, realistic and challenging multimodal database intended for training and testing multi-modal verification systems. The BANCA database was captured in four European languages in two modalities (face and voice). For recording, both high and low quality microphones and cameras were used. The subjects were recorded in three different scenarios: controlled, degraded and adverse over 12 different sessions spanning three months. Each session has two samples. In total 52 people were captured, 26 men and 26 women.

We use one sample from each session for training and the other sample for testing for all the speakers. This procedure is done twice. The overall results are from the average of the two repetitions.

In the recognition, a support vector machine (SVM) classifier was selected since it is well founded in statistical learning theory and has been successfully applied to various object detection tasks in computer vision. Because SVM is defined only for separating two sets of points, the one against one (one-one), and casettoed one against others (one-others) schemes are utilized in our experiments for multi-class classification. For one-one strategy, the SVM models are constructed for each pair of classes. For one-others, each class \(i (i = 1, \ldots, N - 1)\) is split out orderly and all the other classes \(i + 1, \ldots, N\) are merged. The second degree polynomial kernel function is used in the experiments.

Table 1 lists the recognition results using EdgMap-LBP, EdgeMap, LBP and DCT features, respectively. DCT is a commonly used ROI feature [8] for speaker identification. Usually the mouth images need preprocessing, e.g., histogram equalization, for varying lighting conditions across sessions and subjects, or image normalization to eliminate translation, rotation and scaling variations. But our features were extracted from the original detected mouth images without any preprocessing, so the DCT features used in comparison are also extracted from unpreprocessed images. The DCT features are first computed for every frame to get the frame-level features. The final features are obtained by averaging the frame-level features through the segment. The DCT features in the experiment were extracted using square sublattices and five layers of the coefficients selection [19].

From Table 1, it can be seen that with one-one SVM classification, the proposed combined features obtained 82.52% accuracy, higher than only the EdgeMap feature (73.94%) or LBP feature (78.12%). DCT features got only 43.7%. The reason that DCT features do not work is that there is no preprocessing for the affine transformations caused by automatic mouth image localization and illumination variations, but DCT features are very sensitive to these changes. So the results with them are inferior to those with the proposed features.

One-one scheme (second column in Table 1) was compared with casettoed one-others scheme (third column in Table 1) and all the results in Table 1 illustrate the superiority of one-one over one-other strategy. But using this multi two-class strategy, the number of classifiers grows quadratically with the number of classes to recognize. When the class number is \(N\), the number of the SVM classifiers would be \(N(N - 1)/2\). The one-others strategy decompose the \(N\)-class
problem into \( N - 1 \) one-to-rest problems. When the number of speakers is very big, one-others SVM classification is more reasonable and applicable.

### 3.2 XM2VTS Database

We also carried out experiments on the XM2VTS Audio-Visual database [14]. The database consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of the sentence “Joe took fathers green shoe bench out” was used for this research. EdgeMap \_LBP features were extracted from the mouth ROI of the video sequence.

The probe sequences used for testing were obtained from the fourth session. The galleries for training were formed from the first three sessions. Four trials were constructed to test how the performance of the proposed method varied when the time difference between the gallery and the probe set varied between one and three months and multiple sessions were used to form the gallery. For first three trials, there is only one training sample which was captured one (two or three) month ago for each subject. For the fourth trial, there are only two training samples for each subject. It makes the classification very challenging. Because there are only so few training samples for each class, the testing samples are classified or verified according to their difference with respect to the class using the \( k \) nearest neighbor method (\( k = 1 \)). The dissimilarity between a sample and a model EdgeMap \_LBP distribution is measured using the \( L_1 \) distance:

\[
L_1(R_{I}^S, R_{I}^M) = \sum_{n=1}^{N} \left| R_{I}^S(k) - R_{I}^M(k) \right|,
\]

in which, \( R_{I} \) means the EdgeMap \_LBP histogram and \( N \) is the number of bins in the histogram.

The results for DCT features are from [8]. But there are some differences between [8] and our work concerning data used in the experiments: 1) only 251 subjects were used in [8], whereas all 295 subjects are used in our experiments; 2) The mouth ROI was identified manually in [8], whereas it was localized automatically in our work; 3) Many preprocessing steps were used in [8], including that the start and end of some sentences were clipped, mouth ROI was histogram equalized and the mean pixel value was subtracted for dealing with varying lighting conditions across sessions, whereas no such preprocessing was done in our experiment. So our experimental conditions are much more difficult than in [8].

Fig. 6 shows the mouth images from four sessions of the same person (top row) and mouth images from the last session of four different persons (bottom row). Because there is no preprocessing after automatic localization of the mouth area, the translations, rotations and scale variations can be seen from the first row and illumination or skin changes in the second row.

Fig. 7 illustrates the EdgeMap features from XY plane, LBP from XT plane and LBP from YT plane for four samples captured in four sessions for one subject. We can see that for different samples of the same class, their features are very similar to each other, even if they are different in spatial and temporal variation and captured in different time.

>From Table 2, we can see that our method achieved much better results than DCT, EdgeMap, and LBP features. Especially in first and second trials, where the time difference between training sample and test sample is three months and two months, respectively, our results are 21.77\% and 38.97\% higher than that from DCT features. When using multiple sessions in training, the recognition result (75.93\%) is better than using only one training sample.

### 4. DISCUSSION

A novel local spatiotemporal descriptor for speaker identification using sole visual information was proposed, combining dynamic texture features and structural features for motion and appearance representation. The movements of mouth regions are described using local binary patterns in XT and YT planes, for describing the horizontal and vertical motions, while the EdgeMap features extracted from the XY plane are used for describing the structural appearance features. Reliable lip segmentation and tracking is a major problem in speaker identification from mouth movement, especially in poor imaging conditions. Our approach avoids this using local spatiotemporal descriptors computed from mouth regions which are much easier to extract than lips.

Experiments on BANCA database collected from 52 persons using combined features obtained better results than sole EdgeMap features and LBP features, showing the effectiveness of combining features. Speaker identification on the challenging XM2VTS database with 295 subjects show very promising results with over 20\% higher recognition rate than commonly used DCT features.

In the future, it is of interest to combine visual and audio information to promote speaker recognition. We also plan to investigate how to combine visual speech descriptors with face information, with an aim to improve the reliability of face recognition e.g. against spoofing attacks.

### 5. ACKNOWLEDGMENTS

The financial support provided by the Academy of Finland and the EC project IST-214324 MOBIO is gratefully acknowledged.

### 6. REFERENCES


Figure 7: Combined spatiotemporal features for four different samples of same subject. a) EdgeMap features; (b) LBP features from XT plane and (c) LBP features from YT plane.

Table 2: Speaker identification results (%) on the XM2VTS database.

<table>
<thead>
<tr>
<th>Trial #</th>
<th>Gallery</th>
<th>Train</th>
<th>Probe</th>
<th>Test</th>
<th>EdgeMap</th>
<th>LBP</th>
<th>EdgeMap</th>
<th>LBP</th>
<th>DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td>50.85</td>
<td>49.15</td>
<td>40.68</td>
<td>29.08</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td>64.07</td>
<td>55.93</td>
<td>52.54</td>
<td>25.10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td>65.42</td>
<td>58.31</td>
<td>52.88</td>
<td>47.41</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2,3</td>
<td>4</td>
<td></td>
<td></td>
<td>75.93</td>
<td>67.46</td>
<td>64.07</td>
<td>75.30</td>
<td></td>
</tr>
</tbody>
</table>